

Enterprise AI at the Decision Point: A U.S.-Focused Systematic Review of Productivity, ROI, Workforce Change, and Governance

Audrey Rah

arahimi@uh.edu

University of Houston <https://orcid.org/0000-0001-7501-7809>

David W. Johnson

University of Houston

Systematic Review

Keywords: enterprise artificial intelligence, Microsoft 365 Copilot, AI agents, agentic AI, productivity, workforce, ROI, governance, CGEIT, systematic review, United States

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11195857/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Enterprise AI at the Decision Point: A U.S.-Focused Systematic Review of Productivity, ROI, Workforce Change, and Governance

Audrey Rah^{1,*}, David W. Johnson²

¹Department of Electrical and Computer Engineering, University of Houston, Houston, TX, USA

²Assistant Vice President, University of Houston, UIT Technology Services and Support, Houston, TX, USA

*Corresponding author: arahimi@uh.edu

David W. Johnson: dwjohnson@uh.edu

ORCID: Audrey Rah – 0000-0001-7501-7809

Abstract—Enterprise AI assistants and agentic systems have expanded rapidly since 2022, prompting organizations to make increasingly consequential decisions about technology investment, enterprise-wide deployment, productivity expectations, and workforce design. Two claims now dominate much of the public and managerial discussion: that enterprise AI will generate substantial and durable productivity and financial returns, and that it will enable organizations to operate effectively with fewer human workers. Both claims increasingly influence organizational decision making, yet neither has been established consistently across real-world evidence. This systematic review therefore examines what enterprise AI has measurably delivered—particularly in the United States—and whether the available evidence is sufficiently complete to support high-stakes enterprise-governance decisions. ISACA’s CGEIT domains are used as an established governance lens for evaluating decision readiness, particularly for investment, production deployment, scale-up, budget expansion, benefits realization, risk oversight, and workforce restructuring.

Following PRISMA 2020, we reviewed evidence published from January 2022 through 16 September 2026. The search identified 471 records, of which 61 studies were included after screening. The evidence base comprised peer-reviewed and field experiments, U.S. government sources, official institutional evaluations, high-quality analyst research, and clearly labeled vendor materials. Complementary administrative workforce data and targeted searches were used to examine actual AI expenditure, realized ROI, workforce substitution, and the adequacy of evidence available for CGEIT-aligned governance decisions. The findings challenge both dominant claims. U.S. Census data show AI use in 18% of firms (32% employment-weighted). Among AI-using firms, 66% reported augmentation-only use, while 2% reported AI-related employment decreases. Because the latter is self-reported firm attribution, it should not be interpreted as the true rate of AI-driven workforce substitution or as a complete estimate of AI-associated workforce reduction. Causal studies demonstrated meaningful but highly task-dependent productivity effects, including approximately 15% more customer-service issues resolved per hour and 26.08% more completed coding tasks, but also null effects and performance losses in other settings. Thus, observed productivity gains do not generalize uniformly across tasks, tools, or organizational contexts. More importantly for investment decisions, independently measured financial ROI for named enterprise AI assistants was not identified. Among 350 large organizations already using agents, approximately 80% reported some workforce reduction; however, organizations reporting greater workforce reduction did not

consistently report higher ROI. In other words, using AI with fewer human workers did not, by itself, demonstrate stronger financial returns. A targeted substitution analysis identified no complete case ($n = 0$) demonstrating the full evidentiary chain from actual AI cost and budget trade-off to verified workforce replacement and realized ROI. Taken together, the evidence does not support treating either large-scale human-labor substitution or reliably positive enterprise ROI as an established outcome of enterprise AI adoption. The more defensible interpretation is one of uneven augmentation, context-dependent productivity gains, incomplete benefits realization, and major evidence gaps in actual cost, realized ROI, replacement capability, reliability, workforce attribution, and residual risk. These gaps directly constrain CGEIT-aligned executive decisions on whether to move from pilot to production, expand licenses or budgets, scale enterprise deployment, or redesign the workforce around AI. To address this decision problem, the study proposes a minimum evidence package and an evidence-gated sequence for first investment, pilot, production, scale-up, budget expansion, and workforce restructuring.

Index Terms—enterprise artificial intelligence; Microsoft 365 Copilot; AI agents; agentic AI; productivity; workforce; ROI; governance; CGEIT; systematic review; United States

1. INTRODUCTION

SINCE the public release of ChatGPT in November 2022, generative AI has moved rapidly from public experimentation into enterprise products and workflows [1]–[3]. Enterprise offerings now include assistants embedded in productivity software, developer tools, service platforms, and business systems [4]–[6]. Vendors have also moved beyond assistants that draft or summarize toward agents that can perform actions and participate in workflows [7]–[9]. This shift changes the management question. Organizations are no longer deciding only whether employees should use AI. They are deciding whether AI should receive more budget, move from pilot to production, expand across the enterprise, automate workflows, or support work with fewer people. These decisions involve benefits, resources, risk, accountability, and workforce design, which are central concerns of enterprise technology governance [10], [11].

AI adoption is already broad enough for these questions to matter. U.S. evidence shows rapid growth in firm and worker use of generative AI, although adoption varies by firm size, industry, and type of work [12]–[15]. Adoption alone, however, does not establish business value. Two expectations now shape much of the enterprise discussion. The first is that AI will generate substantial productivity and financial returns. The second is that AI will allow organizations to perform the same work with fewer human workers. Existing evidence does not support treating either outcome as automatic. Productivity gains have been demonstrated in specific tasks and settings. Generative AI increased productivity in customer-service work and improved performance in several knowledge-work experiments [16], [17]. Field experiments with software developers also found substantial gains in completed work [18]. Yet other studies show that the effects depend strongly on the task, worker, and technological boundary [19], [20]. AI can therefore improve performance in one setting without producing the same benefit across an organization. The same problem appears in workforce claims. Organizations have reported hiring restraint, restructuring, contractor reductions, and layoffs in connection with AI strategies [21]–[23]. Analyst evidence also shows that workforce reductions do not necessarily correspond to stronger reported returns [24]. A reduction in staffing therefore does not, by itself, demonstrate that AI successfully replaced human capability or created higher financial value. This distinction is important. A statement that many organizations reduced staffing is not a statement that the same percentage of workers was displaced. Organization-level statistics and employee-level displacement measures are different quantities. U.S. labor-market data also provide important context, but aggregate changes in hiring, openings, quits, or layoffs cannot by themselves establish AI causation at a specific firm [25], [26]. Cost and ROI create another evidence problem. Product prices are visible, but list price is not the same as total organizational cost [27]. Enterprise AI can require integration, security controls, governance, monitoring, training, maintenance, and continuing human oversight. Commissioned economic models may estimate returns, but modeled or forecast returns are not equivalent to independently measured realized ROI [28]. This review therefore separates actual expenditure, projected value, and realized financial return.

Security and reliability further complicate scale-up decisions. NIST identifies distinct risks associated with generative AI, including misuse, adversarial behavior, information-integrity problems, and other operational risks [29]–[31]. Agentic systems introduce additional concerns because they may act through tools and external systems rather than only generate text [32], [33]. Documented vulnerabilities such as EchoLeak also show why product capability and production safety must be evaluated separately [34]. These issues make enterprise AI a governance problem as much as a technology problem. ISACA’s CGEIT framework emphasizes governance, IT resources, benefits realization, and risk optimization [10], [35]. COBIT similarly places evaluation, direction, and monitoring at the center of enterprise technology governance through its EDM objectives [11]. NIST, GAO, and U.S. federal guidance also emphasize risk management, accountability, measurement, and

governance throughout the AI lifecycle [29], [36]–[38].

This review uses those governance principles as an analytical lens rather than as evidence that any AI system is successful. It does not classify organizations as CGEIT-compliant or CGEIT-certified. Instead, it asks a practical question: *Does the available evidence give decision makers enough information to justify investment, production deployment, enterprise scale-up, budget expansion, or workforce restructuring?* The core research question is therefore: *What measurable organizational impacts have enterprise AI assistants and AI agents produced, particularly in the United States, across workforce, cost, productivity, ROI, user experience, reliability, security, governance, and operational performance?* Fifteen operational questions (RQ1–RQ15) structure the main evidence extraction. RQ15 maps the findings to the four CGEIT knowledge domains as an analytical scaffold [10]. Additional questions (RQ22–RQ26) examine agent-building difficulty, cybersecurity risk, custom versus low-code implementation, prototype-to-production friction, and hidden human labor.

The main contribution of this review is to connect enterprise AI evidence to the decisions organizations actually have to make. Rather than treating adoption, productivity, workforce reduction, cost, and ROI as interchangeable indicators of success, the review evaluates them as separate links in an evidence chain. This distinction is important because a productivity gain does not by itself establish enterprise ROI, a workforce reduction does not establish successful replacement of human capability, a successful pilot does not establish production readiness, and a projected return does not establish realized financial value.

The review therefore combines six evidence dimensions where the public record permits: (i) enterprise assistants and action-taking agents; (ii) actual organizational AI expenditure rather than list price alone; (iii) mechanism-coded workforce effects; (iv) realized ROI rather than modeled or forecast ROI; (v) hidden human labor and production-readiness evidence; and (vi) a CGEIT/COBIT-aligned evidence-adequacy test [10], [11]. The review also distinguishes framework failure from evidence failure. A governance framework may appropriately require evidence of benefits, resources, risk, and accountability while an organization still lacks the evidence needed to support a sound decision [11], [36]. In such cases, the limitation lies in the available decision evidence rather than necessarily in the governance framework itself. Based on this distinction, the review proposes a minimum evidence package and an evidence-gated sequence for initial investment, pilot, production deployment, enterprise scale-up, budget expansion, and workforce restructuring.

The structure of this paper is as follows: Section 1 introduces the motivation, research problem, objectives, and contributions of the review; Section 2 establishes the enterprise AI context, distinguishing assistive systems from action-taking agents and examining human oversight, expertise, productivity, and enterprise value; Section 3 describes the PRISMA-guided search, screening, evidence classification, data extraction, complementary organizational analyses, and CGEIT/COBIT governance lens; Section 4 presents the findings across adoption, cost and investment, workforce effects, productivity, ROI, user experience, implementation complexity, reliability, security,

governance, and related organizational outcomes; Section 5 interprets the findings and examines the evidence gaps between AI adoption, measurable value, workforce change, and production readiness; Section 6 translates the findings into practical implications for executives, public managers, boards, and assurance functions; Section 7 discusses the methodological and evidence limitations of the review; Section 8 identifies priorities for future empirical research; and Section 9 summarizes the main findings and their implications for evidence-based enterprise AI decision making.

2. BACKGROUND

2.1 Enterprise AI is not consumer AI

Enterprise AI may resemble consumer AI at the interface, but it operates under different organizational conditions. Enterprise systems are embedded in identity, permission, data-protection, logging, and administrative controls that are not equivalent to personal use. Microsoft 365 Copilot operates within Microsoft 365 services and organizational permissions, while ChatGPT Enterprise, Claude Enterprise, and Gemini Enterprise are offered with enterprise-oriented controls and data-handling terms [2], [3], [9], [39]. This distinction matters when interpreting evidence. A controlled experiment may show that generative AI improves a specific task, but it does not by itself establish enterprise ROI, governance, security, or production readiness [17]. This review therefore treats consumer and enterprise use as different evidence contexts.

Fig. 1 summarizes the parallel evolution of enterprise AI adoption, agentic capability, governance challenges, and evidence on human oversight and organizational performance.

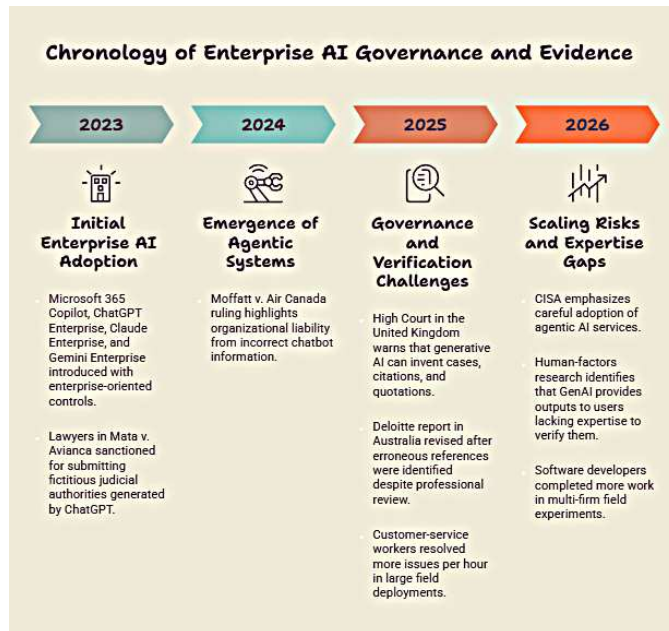


Fig. 1. Chronology of enterprise AI governance and evidence, 2023–2026. The timeline summarizes the progression from initial enterprise AI adoption and early legal accountability cases to the emergence of agentic systems, governance and verification challenges, and later scaling risks related to autonomy, human expertise, and organizational performance.

2.2 The important boundary is what the system can do

For organizations, the important distinction is not the product name but the level of action. Category 1 systems assist. They draft, summarize, search, analyze, recommend, or suggest code. The human remains responsible for the final action. Examples include Microsoft 365 Copilot, ChatGPT Enterprise, GitHub Copilot, Claude Enterprise, and Gemini Enterprise when used for writing, search, coding, or decision support [2]–[4], [40]. Category 2 systems act. They can plan multi-step work, call tools, update records, trigger workflows, or change the state of enterprise systems. Examples include Copilot Studio agents, Salesforce Agentforce, ServiceNow AI Agents, and custom-coded agents [6]–[8]. The same product can fall into either category. A system that drafts a response remains an assistant. The same system becomes agentic when connected tools allow it to modify tickets, CRM records, permissions, or other operational systems. This distinction matters because autonomy changes the consequences of failure. An incorrect answer can often be reviewed before use. An incorrect action may already have changed a system. Agent-security research and NIST evaluations show that prompt injection, tool misuse, and goal hijacking become more important as action capability increases [31], [32], [41]. CISA has similarly emphasized careful adoption of agentic AI services [33].

2.3 Human expertise remains part of the control system

Human oversight is useful only when it can detect error. A reviewer must understand the domain, know what requires verification, and have the authority to reject or escalate an AI output. Real-world cases show that professional expertise alone does not eliminate this risk. In *Mata v. Avianca*, lawyers submitted fictitious judicial authorities generated by ChatGPT and were sanctioned after the citations could not be verified [42]. In the United Kingdom, the High Court later warned that generative AI can invent cases, citations, and quotations and reaffirmed the duty to verify legal authorities [43]. The same issue has appeared at the organizational level. In Australia, a government-commissioned Deloitte report was revised after erroneous references were identified despite professional review [44]. In *Moffatt v. Air Canada*, incorrect information from an airline chatbot resulted in organizational liability [45]. These cases make the governance issue concrete. Human-in-the-loop is not enough if the reviewer lacks time, evidence, expertise, or authority. NIST and GAO both emphasize defined human roles, monitoring, accountability, and oversight in AI governance [29], [30], [36]. For consequential decisions, AI is therefore better treated as decision support than as an unchecked final authority. It can summarize evidence, identify patterns, and generate options, while material decisions remain subject to appropriate human review and intervention [38]. As systems become more autonomous, the need for authorization controls, monitoring, escalation paths, and clearly assigned decision authority also increases [33], [46].

2.4 AI can enable work beyond a user's expertise

Generative AI allows people to perform tasks outside their area of expertise. This expands access to technical and

analytical work, but it also creates an organizational risk. A person may produce a professional-looking output without having the domain knowledge needed to determine whether it is correct. Human-factors research has identified this gap directly: GenAI can provide complete cognitive-task outputs to users who may lack the expertise required to verify them [47]. AI assistance can improve performance without improving self-assessment. Experimental evidence shows that users may perform better while still overestimating the quality of their own work [48]. The result is a gap between production and judgment. AI may increase a person’s ability to produce work faster than that person’s ability to evaluate it. At organizational scale, this gap can become costly. If AI use grows faster than domain expertise, verification, and quality control, visible productivity gains may coexist with hidden rework, technical debt, undetected errors, or weaker decisions. These costs may not appear in task-level productivity measures, but they can reduce the value ultimately realized by the organization. The expertise gap may therefore represent a hidden layer between AI adoption and realized ROI. It should not be treated as a proven cause of weak financial returns, but as a plausible mechanism for why measured productivity gains may not translate into proportional enterprise value. The relevant control is not to restrict AI use to experts. It is to ensure that consequential outputs are validated by people with the expertise required to assess correctness, standards, security, and downstream risk.

2.5 Productivity is not the same as enterprise value

Generative AI can improve specific tasks. Customer-service workers resolved more issues per hour in one large field deployment [16]. Software developers completed more work in multi-firm field experiments [18]. Controlled studies have also reported gains in selected writing and professional tasks [17]. But the effects are uneven. Dell’Acqua and colleagues describe a “jagged technological frontier”: performance can improve when tasks fall within model capability and decline when users rely on the model outside that boundary [19]. Other studies likewise show that benefits vary across workers, tasks, and settings [20]. This matters because a productivity gain is not yet an enterprise outcome. Time saved may become more output, better quality, lower workload, additional capacity, hiring restraint, or workforce reduction. The technology creates capacity; management decides how that capacity is used. Enterprise ROI therefore depends on more than task-level productivity. It also depends on implementation cost, reliability, human oversight, adoption, task mix, quality control, and whether the organization converts the capability into measurable value. The expertise gap adds another layer. An organization may increase the speed or volume of work without increasing its ability to evaluate that work at the same rate. Some apparent productivity gains may therefore coexist with hidden costs that are not captured by conventional performance measures. The same logic applies to agents. Greater autonomy may increase the amount of work that can be automated, but it also increases the importance of monitoring, accountability, authorization, and residual risk [29], [30], [33]. The relevant question is therefore not simply whether AI works. It is whether the organization

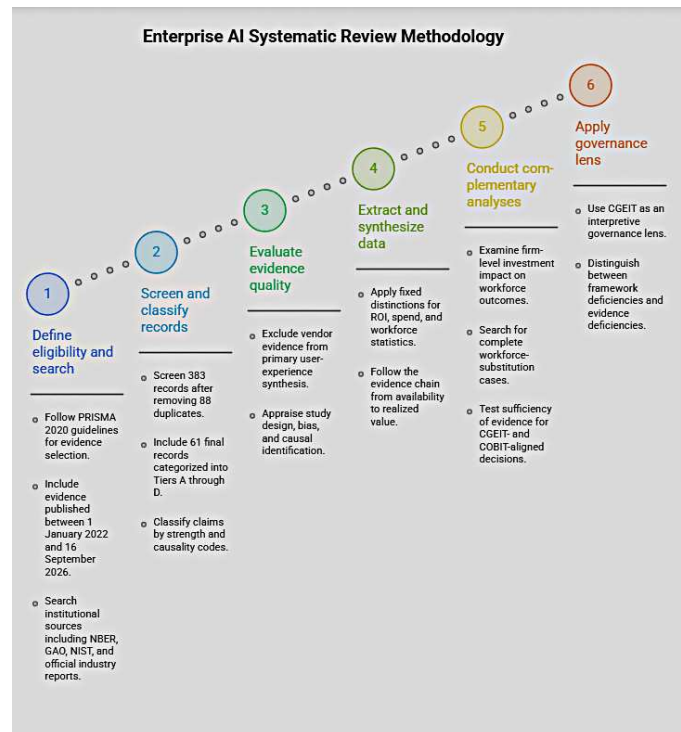


Fig. 2. Enterprise AI systematic review methodology. The workflow progresses from eligibility and evidence identification through screening, evidence-quality assessment, synthesis, complementary organizational analyses, and CGEIT/COBIT-aligned governance interpretation.

can demonstrate measurable value at an acceptable level of risk. That question connects technical performance to enterprise governance. ISACA’s CGEIT domains emphasize benefits realization, resource use, governance, and risk optimization, while COBIT’s EDM objectives emphasize evaluation, direction, and monitoring of enterprise technology decisions [10], [11]. These principles provide the governance lens used later in this review to assess whether the available evidence is sufficient for production deployment, scale-up, budget expansion, and workforce restructuring.

3. METHODS

The review followed a staged methodology from evidence identification and screening to evidence appraisal, synthesis, complementary organizational analyses, and governance interpretation. Fig. 2 summarizes the overall workflow.

3.1 Review design and eligibility

This review follows PRISMA 2020.

Eligible evidence had to meet three conditions: (1) publication or official release between 1 January 2022 and 16 September 2026; (2) evidence on organizational use, deployment, or experimentally relevant use of enterprise AI assistants or agents; and (3) extractable information on methods, sample, geography, and outcomes. Consumer-only chatbot studies, unsourced blogs, SEO pages, and LinkedIn posts were excluded as primary evidence. Major financial and general news sources were used only to document organizational events when

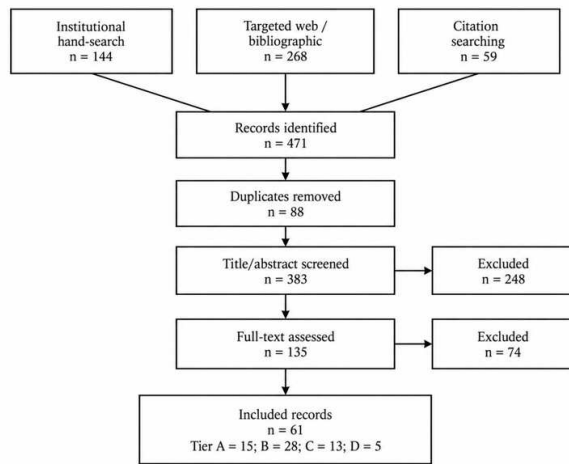


Fig. 3. PRISMA 2020 study-selection flow. The search identified 471 records. After removing 88 duplicates, 383 records were screened, 135 full texts were assessed, and 61 records were included in the final evidence set.

primary or official evidence was unavailable. The review is U.S.-primary, but selected international evidence was retained when it directly informed enterprise deployment, governance, workforce, productivity, or risk.

3.2 Search strategy

The search combined institutional hand-search, targeted bibliographic and web search, citation searching, and DOI verification. Primary sources included NBER, GAO, NIST, the U.S. Census Bureau, BLS, OMB, GSA, GOV.UK, and selected official industry and consulting sources. Peer-reviewed field studies and working papers were also searched directly. Crossref was used to verify publication records and DOIs. Backward citation searching from all included Tier A studies added 41 unique candidates. Scopus and Web of Science were not institutionally available and were not searched. This limitation was mitigated through multi-source institutional searching, backward citation screening, Crossref verification, and targeted bibliographic search. The full search log, screening rules, and verification materials are reported in the Supplementary Materials.

3.3 Screening and study set

The search identified 471 records: 144 from institutional hand-search, 268 from targeted web and bibliographic search, and 59 from citation searching. After removing 88 duplicates, 383 records were screened. Of these, 248 were excluded at title or abstract stage. The remaining 135 full texts were assessed, and 74 were excluded. The final review included 61 records: 15 Tier A, 28 Tier B, 13 Tier C, and 5 Tier D.

Fig. 3 summarizes the study-selection process.

3.4 Evidence strength and claim control

Each major claim was classified by evidence strength. Tier A represents strong independent empirical evidence. Tier B represents authoritative institutional evidence. Tier C represents credible observational or industry evidence. Tier D represents

vendor or self-reported evidence. Claims were also coded as DIRECT, INDIRECT, CLAIMED, or UNKNOWN. Workforce claims received a separate causality code. Direct AI attribution was distinguished from AI contribution, temporal association, non-AI restructuring, no reduction, and unknown attribution. This distinction is central to the review. “AI preceded a workforce change,” “AI contributed to a workforce change,” and “AI caused a workforce change” are treated as different claims. A later link in the evidence chain was never inferred from an earlier one. The evidence chain was:

availability → spending → deployment
→ measured performance → workforce change
→ realized value.

Missing links remained UNKNOWN. No pooled meta-analysis was performed because outcome definitions, samples, and study designs were too heterogeneous for a common effect estimate.

3.5 Independence and quality control

Vendor evidence was not treated as independent proof of usability, ease of use, security, satisfaction, or ROI. For user-experience claims, studies were classified as INDEPENDENT, UNCLEAR, or VENDOR-AFFILIATED using authorship, funding, sponsorship, conflicts of interest, commissioning organization, and data provenance. Only INDEPENDENT evidence entered the primary user-experience synthesis. Quality appraisal considered study design, selection bias, measurement quality, self-report, vendor sponsorship, confounding, causal identification, reproducibility, and generalizability.

3.6 Data extraction and synthesis

A single extraction framework captured organization, country, sector, platform, agent type, deployment setting, sample, cost, users, training, productivity, time saved, ROI, workforce change, user experience, reliability, security, oversight, limitations, evidence tier, causality, and independence. Several distinctions were fixed before synthesis. List price was not treated as actual organizational spend. Modeled or commissioned ROI was not treated as realized ROI. Organization-level workforce statistics were not treated as employee-level displacement. Reported AI use was not treated as evidence of AI expenditure. A task-level productivity gain was not treated as evidence of realized enterprise value. These rules were applied consistently across the review.

3.7 Complementary organizational analyses

Five complementary analyses extended the systematic review without changing the 61-record PRISMA set. First, we examined whether greater firm-level AI investment was followed by changes in hiring, headcount, attrition, or other workforce outcomes. Second, we examined U.S. administrative labor-market indicators, while keeping aggregate industry evidence separate from named-firm evidence. Third, we searched for complete workforce-substitution cases. A complete case required evidence of actual AI expenditure, budget reallocation,

verified workforce action, demonstrated replacement capability, and realized financial return. Fourth, we searched across major enterprise platforms for actual organizational AI expenditure. Vendor list prices and guessed seat counts were not treated as spend. Fifth, we tested whether the public evidence available for enterprise AI was sufficient for CGEIT- and COBIT-aligned decisions on scale-up, budget allocation, production deployment, and workforce restructuring. These analyses followed one rule: missing evidence was not inferred.

3.8 CGEIT as a governance lens

CGEIT was used as an interpretive governance lens, not as a compliance score. RQ15 maps evidence to CGEIT domains using qualitative labels such as supported, partly supported, weak, or not reported. COBIT EDM01–EDM05, NIST AI RMF, the GAO AI Accountability Framework, and OMB M-25-21 were used to identify the evidence that a decision packet should contain [11], [36], [38]. The review therefore distinguishes two different problems. A framework deficiency means the governance framework does not require an important decision input. An evidence deficiency means the framework requires the input, but the available enterprise-AI evidence does not provide it. This distinction prevents evidence gaps from being misinterpreted as failures of the governance framework.

4. RESULTS

The evidence base is uneven. Strong causal organizational evidence is concentrated in a small number of field experiments, while nationally representative U.S. statistics provide the broadest firm-level picture. In contrast, major commercial agent platforms are largely absent from Tier A organizational outcome studies.

4.1 Characteristics of included evidence

Table I summarizes the included set. Causal organizational evidence is concentrated in a small number of field experiments: the Dillon et al. Microsoft 365 Copilot RCT [49] (66 firms; 7,137 workers); the Brynjolfsson et al. customer-service study [16] (5,172 agents); the Cui et al. three-firm developer RCT [18] (4,867 developers); the Dell’Acqua BCG and P&G experiments; the Afshar et al. U.S. ambient-scribe RCT [50] (66 clinicians); and several non-enterprise or non-U.S. experiments retained as comparators. U.S. official statistics provide the only nationally representative firm picture. Named commercial agent platforms are almost absent from Tier A organizational outcome studies.

4.2 RQ1 — Adoption

RQ1 examines how widely enterprise AI is actually being adopted. Reported adoption rates vary because firm use, worker use, and executive survey responses measure different things and should not be directly compared. U.S. official firm data and global executive surveys describe different phenomena and must be separated. Using the post-17 November 2025 BTOS wording (AI in any business function), Bonney and colleagues [14] report that 18% of firms used AI during November 2025–January 2026, rising to 32% on an employment-weighted basis,

with expected use of 22% within six months. Worker task use was reported by 23% of firms (41% employment-weighted). The Census Bureau’s May 2026 story covering 14 December 2025–3 May 2026 finds overall usage hovering between 17% and 20%, with 37% of firms with at least 250 employees using AI and 32% of firms with 100–249 employees in the period ending 3 May 2026. A Federal Reserve FEDS Note [51] (3 April 2026) underscores unit-of-analysis gaps: about 18% of firms (BTOS, end-2025), about 41% of individuals reporting work-related generative AI (Real-Time Population Survey, November), and 78% of the labor force employed at firms that have adopted AI (Survey of Business Uncertainty).

These official rates are far below the McKinsey 2025 finding [52] that 88% of survey respondents regularly use AI in at least one function (versus 78% a year earlier), with about one-third beginning enterprise-wide scaling and 23% scaling agents somewhere (39% experimenting). PwC’s 2025 global worker survey [53] (n = 49,843) finds that 54% used AI for their job in the past year, but only 14% used generative AI daily and 6% used agentic AI daily. Bick, Blandin, and Deming’s household evidence [12] (February 2025 revision) reports nearly 40% population use of generative AI and 23% of employed adults using it for work in the prior week, with time savings of 1.4% of work hours; a later Management Science / St. Louis Fed version reports higher use rates (45% / 27%). We cite the version explicitly and do not blend them. Fig. 4 displays the incomparability.

Federal generative AI use cases in 11 inventoried agencies rose from 32 in 2023 to 282 in 2024 [37], while total AI use cases rose from 571 to 1,110. GSA OneGov announcements in August–September 2025 created promotional access to ChatGPT Enterprise (including a \$1-per-agency year in the 6 August 2025 release) and no-cost Microsoft 365 Copilot for up to 12 months for qualifying Microsoft G5 customers (2 September 2025). Seat counts are not published in those releases. ChatGPT Enterprise telemetry [54] (2026 working paper; OpenAI-affiliated) shows rapid within-firm growth in tokens through March 2026 but does not measure productivity. Peer-reviewed organizational outcome studies of Salesforce Agentforce, ServiceNow agents, Claude Enterprise, and Gemini Enterprise were not verified in this search. Table II records extractable scale.

4.2.1 Commercialization chronology (not adoption): A product announcement is not organizational adoption, and a public list price is not organizational spending. Table III is a U.S.-relevant commercialization chain from consumer generative AI (November 2022) through paid enterprise assistants (ChatGPT Enterprise, 28 August 2023; Microsoft 365 Copilot GA, 1 November 2023) to vendor-claimed agent platforms (Agentforce GA, 29 October 2024; Gemini Enterprise, 9 October 2025) and federal promotional procurement (GSA OneGov, August–September 2025). GitHub Copilot’s June 2022 business availability precedes ChatGPT and is a paid coding-assistant SKU, not proof of enterprise-wide deployment. Microsoft’s September 2023 statement that tens of thousands of EAP users already used Copilot is a vendor claim, not an invoice census (Microsoft 365 Blog, 21 September 2023). Census BTOS remains the only nationally representative U.S. firm picture:

TABLE I

SELECTED INCLUDED RECORDS. FULL EXTRACTION (S001–S061) IS IN ENTERPRISE_AI_EVIDENCE_MASTER.XLSX. TIERS: A INDEPENDENT EMPIRICAL; B INSTITUTIONAL; C OBSERVATIONAL/INDUSTRY; D VENDOR/SELF-REPORT. S028 IS SHOWN FOR SCALE ONLY AND IS EXCLUDED FROM INDEPENDENT UX SYNTHESIS.

ID	Setting	Technology	Design (n)	Tier	Geo
S001	66 firms	M365 Copilot	RCT; 7,137 workers	A	US/EU mix
S002	F500 software CS	GPT reply assistant	Staggered DiD; 5,172 agents	A	US customers; mostly PH agents
S003	MSFT, Accenture, F100	Coding assistant	3 RCTs; 4,867 developers	A	US-HQ firms
S004	Upwork task	GitHub Copilot	RCT; 95/35 completers	A	Intl freelance
S005	Prolific writers	ChatGPT 3.5	RCT; 453	A	Online
S006	BCG	GPT-4	RCT; 758 consultants	A	Intl
S007	P&G	Generative AI	Field exp.; 791	A	US HQ
S009	WI/IL clinics	Abridge in Epic	Stepped-wedge; 66	A	US
S014	US firms	Any business-function AI	BTOS supplement	B	US
S018	12 federal agencies	GenAI inventories	GAO audit	B	US
S027	UK government	M365 Copilot	20,000 licenses; survey 7,115	B	UK
S028	Virginia Tech	M365 Copilot	Pilot 177; survey 85 (vendor-partner; UX-excluded)	B	US
S049	UK DWP	M365 Copilot	3,549 licences; surveys 1,716/2,535	B	UK
S050	UK DBT	M365 Copilot	1,000 licences; diary/interviews	B	UK
S051	UK HMRC	M365 Copilot	3,000 random licences; survey 1,364	B	UK
S032	Large enterprises	Agents/automation	Gartner n=350	C	Global
S041	Composite firm	M365 Copilot	Forrester TEI (commissioned) [28]	D	Modeled

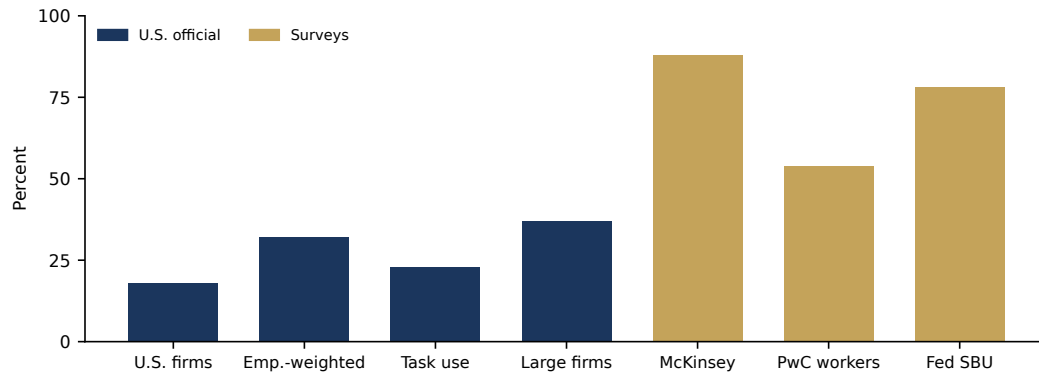


Fig. 4. Incomparable adoption measures for generative or business-function AI, 2024–2026. Bars report the statistic as published: U.S. firm share, employment-weighted firm share, worker task-use share, and large-firm share (CES-WP-26-25; Census BTOS story, May 2026); McKinsey 2025 respondent share of functional AI use; PwC 2025 worker share of job-related AI use; and the FEDS Note (3 April 2026) unit-of-analysis series. These quantities must not be averaged.

TABLE II

NAMED PLATFORMS WITH EXTRACTABLE SCALE (SECTION 4.2). “NONE VERIFIED” = NO PEER-REVIEWED ORGANIZATIONAL-OUTCOME RCT LOCATED FOR THAT COMMERCIAL AGENT PLATFORM.

Platform / context	What is known about scale	Source	Tier
Any AI, US firms	18% of firms; 32% emp.-weighted (Nov 2025–Jan 2026)	[14]	B
M365 Copilot RCT	7,137 workers; 66 firms; 300-license incentive	Dillon et al. 2025 [49]	A
M365 Copilot, UK gov	20,000 licenses; 12 organisations	[55]	B (intl)
M365 Copilot, Virginia Tech	177 volunteer seats; 38 workdays	VT 2025	B
M365 Copilot, CSIRO	300 licenses; 6-month trial	Bano et al. 2024 [56]	C (intl)
ChatGPT Enterprise	>1,500 orgs in telemetry WP; seats unpublished in WP	Chatterji et al. 2026 [54]	B/D mix
Federal GenAI use cases	32 → 282 (2023–24), 11 agencies	[37]	B
Named agent-platform org RCTs	None verified	This search	—

18% of firms used AI in any business function in November 2025–January 2026 [14]. Dynamics 365 Copilot (announced

TABLE III
ENTERPRISE AI COMMERCIALIZATION AND SPENDING-TYPE TIMELINE, 2022–16 SEPTEMBER 2026. FULL ROWS ARE IN
ENTERPRISE_AI_LONGITUDINAL_EVIDENCE.XLSX. ANNOUNCEMENT $\neq$ ADOPTION. LIST PRICE $\neq$ EXPENDITURE.

Date	Technology	Event / stage	Spend type	Source
2022-06-21	GitHub Copilot	Paid SKU GA — not org census	Subscription list	GitHub; S004
2022-11-30	ChatGPT consumer	Consumer release — not enterprise tenancy	None	OpenAI
2023-08-28	ChatGPT Enterprise	Enterprise GA; price not public list	Usage / negotiated	OpenAI; Reuters
2023-11-01	M365 Copilot	Enterprise GA; \$30 list; 300-seat min. later dropped	List, not invoice	Microsoft Tech Community
2023-11-15	Copilot Studio	Low-code agents announced (Ignite)	Capacity list, not TCO	Microsoft Ignite
2024-02-27	GitHub Copilot Enterprise	Enterprise SKU GA	List/subscription	GitHub
2024-01-16	M365 Copilot	No seat minimum; more base SKUs	List \$30	Microsoft partner FAQ
2024-10-29	Agentforce	GA; vendor-named users unverified here	List/credits, not invoice	Salesforce IR
2025-08/09	Federal OneGov	Promo ChatGPT \$1/agency; Copilot \$0 up to 12 mo G5	Committed promo; \$3.1B claimed	GSA (S025–S026)
FY2025	Microsoft PPE	Vendor capex, not buyer Copilot TCO	Actual 10-K \$64.551B additions	MSFT 10-K FY2025
2025-10-09	Gemini Enterprise	Cloud workplace/agent platform launch	Press list \$30/\$21	Google; Reuters
2025-11–2026-01	Any firm AI	18% of U.S. firms; 32% emp.-weighted	Not spend	[14]

6 March 2023), Copilot Studio (Ignite, 15 November 2023), GitHub Copilot Enterprise (GA 27 February 2024), Claude Team/Enterprise tenancies (2024–25), ServiceNow Now Assist and AI Agents (2023–25 product waves), and Gemini Enterprise (9 October 2025) enter the chronology as product-availability events; none of those dates is treated as organizational adoption or as independently verified agentic scale-up. Fig. 12 places these dates on Track A without drawing a causal arrow to labor outcomes. Section 4.13.1 and Fig. 14 test which later links (procurement, use, productivity, hiring, layoff explanation) are actually evidenced. Copilot Studio, Agentforce, ServiceNow AI Agents, and Gemini Enterprise enter as action-capable product families; availability is not production proof (Section 2.2).

4.3 RQ2 — Cost and investment

RQ2 asks a simple question: what does enterprise AI actually cost organizations, and how much of the reported financial value is real rather than listed, modeled, or promotional? Vendor list price and organizational expenditure are different objects (Section 3.4). Microsoft’s public enterprise page, accessed 16 September 2026, lists Microsoft 365 Copilot at \$30.00 per user per month paid yearly, requiring a separate qualifying Microsoft 365 license; the page states that displayed prices are for marketing purposes. Copilot Studio capacity is sold separately in vendor commercial materials (not independently audited here). Google’s Gemini Enterprise list around the October 2025 launch was reported at \$30 per seat per month for standard enterprise and \$21 for a business tier (press); we treat those as vendor list, not invoices.

U.S. federal promotional prices diverge sharply from commercial list: ChatGPT Enterprise at \$1 per agency for one year (GSA, 6 August 2025) and Copilot at no cost for up to 12 months for G5 customers (GSA, 2 September 2025). GSA’s claim of \$3.1 billion in potential first-year Microsoft-related savings is a government promotional figure, not an

independently audited realization, and is coded CLAIMED. The Forrester TEI study [28] models a 25,000-employee composite with \$36.8 million in three-year benefits, \$17.1 million in costs, \$19.7 million NPV, 116% ROI, and 10-month payback, based on 16 interviews across 12 organizations and a survey of 367 users. Forrester [28] states that it makes no assumption that other organizations will receive this ROI. We code the 116% figure as modeled and vendor-commissioned (Tier D), not independently measured. IDC’s 2025 press forecast of AI IT spend reaching \$1.3 trillion in 2029 is a market forecast, not a cost study of assistants. Table IV distinguishes list, promotional, modeled, and claimed figures.

When organizations began paying meaningful amounts is only partly observable. Actual expenditure that can be verified in this search is concentrated in (a) vendor capital expenditure to supply AI infrastructure—Microsoft’s Form 10-K for the fiscal year ended 30 June 2025 reports additions to property and equipment of \$64.551 billion, versus \$44.477 billion in FY2024 and \$28.107 billion in FY2023, and states that capital expenditures will continue to support cloud offerings and AI infrastructure (Microsoft 10-K, cash-flow statement and MD&A)—and (b) public promotional federal prices, which are committed contractual terms, not commercial invoices (GSA, 6 August and 2 September 2025). Those vendor capex figures are not customer Copilot TCO and must not be averaged with the \$30 list add-on or with the Forrester modeled \$17.1 million composite cost (S041) [28]. Organizational license invoices, implementation, training, and security budgets are almost never published in 10-K line items as “enterprise assistant spend.” Planned or CEO-projected workforce savings (IBM, May 2023) are not spending. Table XV records that no included organization has a complete pair of documented AI invoices and independently verified subsequent layoffs (Section 4.13; attribution coding in Section 4.13.1).

TABLE IV
COST AND INVESTMENT EVIDENCE EXTRACTED FROM INCLUDED SOURCES. LIST, PROMOTIONAL, MODELED, AND CLAIMED FIGURES ARE DIFFERENT OBJECTS AND ARE NOT INTERCHANGEABLE.

Item	Amount	Type	Source	Tier
M365 Copilot enterprise add-on	\$30 user/month yearly	US list (marketing page)	Microsoft 2026	D
Qualifying M365 base license	Additional; not in \$30	List / contract	Microsoft 2026	D
ChatGPT Enterprise (OneGov promo)	\$1 per agency / 1 year	Promotional federal	GSA 2025-08-06	B
M365 Copilot (OneGov G5 promo)	No cost up to 12 months	Promotional federal	GSA 2025-09-02	B
GSA claimed Microsoft savings	\$3.1B potential year 1	CLAIMED	GSA 2025-09-02	B/ CLAIMED
Forrester TEI composite 3-year cost	\$17.1M modeled	Modeled TCO	Forrester TEI 2025 [28]	D
Forrester TEI composite 3-year benefit	\$36.8M modeled	Modeled	Forrester TEI 2025 [28]	D
NANDA investment context	\$30–40B enterprise GenAI	Author context, not a ledger	Challapally et al. 2025 [57]	C

4.4 RQ3 — Workforce effects

RQ3 examines whether organizations are actually replacing workers with AI, reducing hiring, or changing employees' roles because of AI. The U.S. nationally representative firm evidence does not show mass AI-caused layoffs through early 2026. Bonney et al. [14] find that 66% of AI-using firms use AI only to augment tasks and that AI-related employment decreases occur in only 2% of firms. That 2% is a firm share, not an employment-weighted job-loss rate, and is self-reported. BLS 2024–34 projections incorporate assumed AI productivity effects that raise some technology occupations (software developers +15.8%, or +267,000 jobs; data scientists +33.5%) and dampen some office and customer-service occupations; these are projections, not measured enterprise-assistant impacts (causality code 0).

Analyst surveys, which over-represent large firms already using automation, show more reduction activity—but still as organization shares. Gartner (5 May 2026) surveyed 350 global executives at firms with at least \$1 billion in revenue that were piloting or had deployed AI agents, intelligent automation, or autonomous technologies (Q3 2025). Approximately 80% reported workforce reductions. Reduction rates were nearly equal among respondents reporting higher ROI and those reporting modest or negative outcomes. Helen Poitevin's quoted interpretation is that reductions may create budget room but do not create return. We code this as causality 2 at the organization-attribution level and emphasize that 80% is not a percentage of employees. A Computerworld paraphrase that some cuts reached 20% of headcount does not appear in the official Gartner press-release text we verified and is not attributed to Gartner in this review.

In customer service, Gartner's October 2025 survey of 321 leaders (released 3 February 2026) found that only 20% had actually reduced agent staffing due to AI, while a majority reported steady headcount while serving more customers. The same release's 50% "rehire by 2027" statement is a Gartner prediction, not a survey percentage. The 28 April 2026 release, using the same $n = 321$ fielded September–October 2025, reports that 85% are expanding human-agent responsibilities, 75% are shifting agents into new roles, 63% are reducing frontline headcount through attrition, and 31%

have implemented or are planning AI-driven frontline layoffs through the first quarter of 2027. The 31% mixes implemented and planned actions. Gartner's July 2026 CHRO survey [23] ($n = 110$, 4Q25) finds that 22% of CHROs report that at least one business leader stopped hiring for entry-level roles due to AI automation; 95% of those organizations had implemented AI in some capacity, but only one in five realized significant or transformational value. "At least one leader" is not an organization-wide freeze and not a 22% decline in junior employment.

PwC's April 2025 U.S. executive agent survey [58] ($n = 308$) finds that 48% say they will likely increase headcount and 67% agree agents will drastically transform roles within 12 months—an expectation series, not payroll. The IRS case [59] is a critical counterexample: the Research, Applied Analytics and Statistics group lost 63 employees who had supported AI (about 10% of January 2025 RAAS staffing) amid 2025 staffing reductions. GAO does not attribute those losses to AI substitution; officials still planned to use more AI, and GAO warned that the absence of a skills-gap plan raises the risk that AI efforts will not succeed. Causality code 4. Klarna's February 2024 press release [60] claimed an assistant handled two-thirds of customer-service chats, equivalent to the work of 700 full-time agents, with an estimated \$40 million 2024 profit improvement. Those 700 are workload equivalents, not verified layoffs. Bloomberg reporting in May 2025 described a subsequent quality-driven return to human customer service after a hiring freeze of more than a year. We code Klarna as international, vendor/news evidence, causality 2, and do not treat 700 as a U.S. layoff count. Fig. 5 and Table V display these findings as organization or respondent shares, not as percentages of employees.

4.4.1 Longitudinal labor demand and hiring (U.S.): Job-posting declines after 2022 are not automatically AI-caused. The 2022–23 technology-sector layoff wave largely preceded Microsoft 365 Copilot general availability (1 November 2023) and coincided with post-pandemic hiring normalization and higher interest rates; those events are coded temporal association / macroeconomic unless a filing names AI as the mechanism. Against that backdrop, two academic series are specific enough to report with denominators.

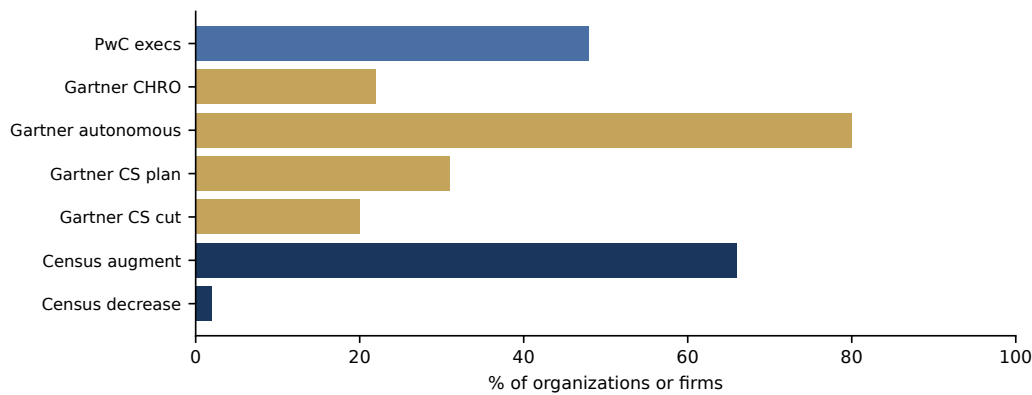


Fig. 5. Workforce-related outcomes reported as organization or respondent shares, not as shares of employees laid off. Sources: CES-WP-26-25 (U.S. firms); Gartner newsroom releases of 3 February, 28 April, 5 May, and 27 July 2026; PwC U.S. Agent Survey 2025 (expectations). Causality codes follow Section 3.4.

TABLE V

WORKFORCE EFFECTS IN THE INCLUDED SET. PERCENTAGES ARE ORGANIZATION OR RESPONDENT SHARES UNLESS A SOURCE STATES OTHERWISE. CODES: 1 DIRECT AI ATTRIBUTION; 2 AI CONTRIBUTING; 3 TEMPORAL ONLY; 4 PRIMARILY NON-AI; 5 NO REDUCTION; 0 UNKNOWN.

Finding	Unit	n / frame	Code	Tier
AI-related employment decrease in 2% of firms; 66% of users augment-only	US firms	BTOS supplement	2 / 5	B
~80% of large agent-using orgs report some reduction; no ROI correlation	Global \$1B+ orgs	350 execs, Q3 2025	2	C
20% of CS leaders reduced agent staffing due to AI	CS orgs	321 leaders, Oct 2025	2	C
31% implemented or planning AI-driven CS layoffs through 1Q27	CS orgs	321, Sep–Oct 2025	2	C
85% expanding human CS responsibilities; 75% new roles; 63% attrition path	CS orgs	321	5/2 mix	C
22% of CHROs: ≥1 leader stopped some entry-level hiring due to AI	HR heads	110, 4Q25	2	C
IRS RAAS lost 63 AI-supporting staff	One federal unit	[59]	4	B
Klarna 700 FTE-equivalent (PR); later human reinvestment (news)	One firm, intl	PR + Bloomberg	2	D

Westby, Modestino, and Cheng [61] (IZA Discussion Paper 18723, June 2026) use Lightcast U.S. software-developer vacancies (SOC 15-1252). In an event-study / difference-in-differences window around ChatGPT’s November 2022 release, junior postings requiring three or fewer years of experience fell roughly 14–15 percent relative to postings requiring four or more years. Descriptively, from October 2022 to October 2024, junior vacancies that required at least one AI/ML skill fell from 1,915 to 1,029 (–46.3%), while senior AI/ML vacancies fell from 4,338 to 3,934 (–9.3%), amid a broader software-posting contraction. Attribution: AI-contributing evidence for relative junior software demand, with remaining confounding from sector contraction. This is not a named Copilot or Agentforce effect.

Brynjolfsson, Chandar, and Chen [62] (Stanford Digital Economy Lab, November 2025) use ADP payrolls. After firm-time controls, employment of workers aged 22–25 in the most AI-exposed occupations was about 16 percent lower than that of experienced workers in the same firms through September 2025; adjustments were concentrated where AI is used to automate rather than augment, and results are reported as robust to dropping technology firms. Attribution: AI-contributing (occupation-exposure design). It is not a count of Copilot-caused layoffs and is not an economy-wide employment collapse. Acemoglu, Autor, Hazell, and Restrepo [63] (NBER w28257) show that even in 2010–18, AI-exposed establishments reshuffled vacancies toward AI skills without

detectable occupation- or industry-level employment effects—a caution against reading posting mix as headcount.

Official U.S. firm data still show rare self-reported AI-related employment decreases (2% of firms) and modal augmentation [14]. Gartner’s 22% of CHROs (n = 110, 4Q25) reporting that at least one leader stopped some entry-level hiring due to AI must not be rewritten as a 22% freeze of junior employment (Section 4.4). Table VI, Table VII, and Table VIII record baselines, populations, attribution, and named layoff or restructuring cases. Fig. 12 is Track B. Occupation-complete posting series for customer service, administrative support, HR, marketing, and finance were not extracted as a monthly Lightcast panel in this package; those cells are coded insufficient rather than imputed from software vacancies (Section 4.13.1).

4.4.2 Indirect and non-explicit AI-related workforce change: Many organizations never write “AI caused layoffs.” This subsection therefore does not stop at that phrase. It asks whether a documented mechanism connects AI investment or use to a workforce action: investment/deployment → efficiency or operating-model program → hiring freeze or non-backfill → task automation → smaller headcount or planned growth → management explanation → any budget reallocation. Hidden forms include attrition without replacement, reduced entry-level hiring, role consolidation, contractor non-renewal, and higher output expectations with fewer people. Alternative explanations (revenue, overhiring, mergers, rates, outsourcing,

TABLE VI
U.S. JOB-POSTING AND HIRING INDICATORS. FULL FIELDS: ENTERPRISE_AI_LONGITUDINAL_EVIDENCE.XLSX. POSTING DECLINES ARE NOT AUTOMATICALLY AI-CAUSED.

Indicator	Window	Population	Estimate	Attribution
JOLTS job openings (all industries)	Peak Mar 2022 → later years	U.S. nonfarm openings	Post-peak decline (BLS levels)	Macro / post-pandemic — not AI
Junior vs senior software vacancies	Around Nov 2022; Lightcast	U.S. SOC 15-1252	~14–15% relative junior decline	AI-contributing; sector confounders
Junior AI/ML software vacancies	Oct 2022–Oct 2024	Counts 1,915 → 1,029	-46.3% junior vs -9.3% senior	Descriptive + market contraction
Ages 22–25 in most AI-exposed occ.	Late 2022–Sep 2025 ADP	Payroll; firm-time controls	16% relative employment decline	AI-contributing; not product-causal
U.S. firms AI-related emp. decrease	Nov 2025–Jan 2026	BTOS firms	2% of firms; 66% augment-only	Official self-report; not worker share
CHRO: ≥1 leader stopped some junior hiring	4Q25 / Jul 2026	110 CHROs	22% of CHROs	Survey; not org-wide freeze
Tech layoffs 2022–23	Before Copilot GA	Several U.S. tech firms	Firm-specific	Macro / overhire; temporal only

TABLE VII
HIRING FREEZES AND STAFFING REDUCTIONS. EXACT WORDING AND SOURCES ARE IN THE LONGITUDINAL WORKBOOK. “AT LEAST ONE MANAGER” ≠ ORGANIZATION-WIDE FREEZE.

Organization	Date	Action	AI attribution
IBM	2023-05-01	Pause/slow back-office hiring; up to ~7,800 over 5 years via AI/automation incl. attrition (projection)	AI-contributing plan; not a completed layoff census
Gartner CHRO sample	4Q25	22% : ≥1 leader stopped some entry-level hiring due to AI	Survey; not one-firm freeze
Gartner CS sample	Sep–Oct 2025	20% reduced agent staffing due to AI	Survey
Klarna (intl)	2024–25	700 FTE-eq claim; later human CS return	Vendor then reversal

TABLE VIII
LAYOFF AND RESTRUCTURING CASES. DIRECT AI SUBSTITUTION IS NOT ESTABLISHED FOR A U.S. NAMED ENTERPRISE-ASSISTANT LAYOFF WAVE. FULL EXTRACTION IN THE WORKBOOK.

Case	Labor outcome	Causality class
Census firms	2% report AI-related employment decrease	Unknown at worker level
IRS RAAS	63 AI-supporting staff lost (2025)	General restructuring (code 4)
IBM	Planned back-office reduction / hiring pause	AI-contributing planned
Gartner n=350	~80% of agent-using orgs some reduction; ROI bands similar	AI-contributing survey
2022–23 tech wave	Large firm-specific cuts	Temporal / macroeconomic
Klarna	Claimed 700 FTE-eq; 2025 human return	AI-contributing then reversal
Dillon RCT firms	No layoff outcome reported	No reduction finding

sector contraction) are tested first. Stronger classes are kept only when the AI mechanism still stands after those alternatives.

Seven classes are used (they refine, and do not silently replace, codes 0–5 / A–F). 1 Explicit AI substitution — the organization states that AI/automation replaces specified work or roles. 2 AI-enabled workforce compression — headcount or planned growth falls while management explicitly expects AI/automation to preserve or raise output. 3 AI-prioritized budget reallocation — people or other spend is cut while money is explicitly redirected to AI. 4 AI-related non-backfill / hiring suppression — vacated or planned roles are not filled because AI/automation is expected to absorb the work. 5 AI-contributing restructuring — AI is one documented factor among several. 6

General restructuring with AI present — both occur in the same period with no connecting mechanism. 7 Temporal association only. For classes 2–5 this review does not write “AI caused the layoff.” It writes: AI appears to have materially contributed to the workforce decision through the documented mechanism of [productivity substitution / non-backfill / budget reallocation / operating-model redesign]. If the mechanism is incomplete, the case is coded lower.

Named U.S.-relevant dossiers. IBM (1 May 2023, CEO interview via Bloomberg/Reuters): pause or slow back-office hiring; up to ~7,800 roles over five years as 30% of ~26,000 non-customer-facing, including attrition that may not be backfilled. Class 4, planned. Missing: named SKU, invoices, completed

TABLE IX

AI-WORKFORCE ATTRIBUTION DOSSIERS. FULL COLUMNS ARE IN ENTERPRISE_AI_LONGITUDINAL_EVIDENCE.XLSX. CLASSES 2–5 USE CONTRIBUTION LANGUAGE, NOT 'AI CAUSED THE LAYOFF.'

Organization	Workforce action	Class	Mechanism (not 'AI caused')	Alternative
IBM, May 2023	Hiring pause; planned attrition non-backfill	4 planned	Non-backfill of back-office roles AI 'could' do	Vendor-user strategy; earlier cuts
Salesforce, Feb 2025 call	No new engineer hiring	4 + 2 language	Claimed +30% eng. productivity; mix toward sales	Margin program; vendor interest
Salesforce 2023–24 cuts	Large reductions	6 / 7	None established to Agentforce	Overhire / cost
Duolingo, 2023–24	~10% contractors not renewed	5 (contractor)	AI part of content-ops; not FTE layoff	Contract end; ops redesign
Bank of America calls	Ops roles down via attrition; client-facing up	5 + 2/4 language	AI/digital work elimination	Pre-GenAI digitalization
IRS RAAS	63 AI-supporting staff lost	6	None — opposite of substitution	2025 federal cuts
Tech wave 2022–23	Firm-specific layoffs	7 / 6	None to Copilot invoices	Rates; overhire

7,800-person layoff. Salesforce FY2025 Q4 earnings remarks (February 2025): no new software-engineer hiring, citing ~30% engineering productivity from AI coding tools and Agentforce; sales hiring to rise. Class 4 with class 2 language. Do not write that AI caused a 2025 engineer layoff; prior 2023–24 Salesforce reductions remain class 6/7. Duolingo (company comment to Bloomberg, January 2024): ~10% of contractors not renewed; 'part of that could be attributed to AI'; no full-time employees affected. Class 5 with class 1 language for rote contractor tasks. Bank of America earnings-call series: AI/digitalization used to eliminate operational work, add client-facing staff, and hold or reduce headcount via attrition. Class 5 with class 2/4 language; long-run digitalization is an alternative that predates ChatGPT. IRS RAAS remains class 6 (GAO). The 2022–23 technology layoff wave remains class 7/6. No named U.S. dossier in this search is a completed, independently verified class 1 employee-layoff wave tied to a named enterprise-agent invoice. Class 3 (explicit payroll-to-AI budget vote) was not identified as a complete ledger.

Frequency. In the available named-organization public record, indirect forms—hiring suppression, contractor non-renewal, attrition-with-output language—appear more often than completed explicit 'AI layoff' counts. The largest absolute 2022–23 headcount reductions remain class 6/7 events. Survey items (Gartner CHRO 22%; CS 20% staffing / 63% attrition path) are consistent with compression and non-backfill as the common self-reported path, but they are organization shares, not a WARN census. Table IX is the dossier.

4.5 RQ4 — Productivity and time savings

RQ4 examines whether AI improves productivity, saves time, or improves work quality, and whether these task-level gains translate into organization-wide performance and realized enterprise ROI. Causal productivity evidence is real, bounded, and task-specific. In the QJE version of Generative AI at Work, access to a GPT-family suggestion tool increased issues resolved per hour by 15% among 5,172 customer-support agents, reduced average handle time by 3.7 minutes (8.5% from a 43-minute baseline), and increased chats handled per hour by 0.37 from 2.4. Agents followed suggestions 35% of the time. Gains were larger for less experienced workers; the most experienced saw small speed gains and small quality declines. Eleven percent of agents were in the United States.

Cui et al. [18] (Management Science, 2026), in three firm RCTs totaling 4,867 developers, report a combined 26.08%

increase in completed tasks (standard error 10.3%) among developers using an AI coding assistant, with larger adoption and gains among less experienced developers. Site-level results were noisy. Peng et al. [64] found that completers of a short JavaScript server task finished 55.8% faster with GitHub Copilot (71.17 versus 160.89 minutes; $p = 0.0017$; 95% CI 21–89%), but only 35 of 95 accepted participants completed the task, and the setting is not an enterprise deployment. Noy and Zhang [17] found 40% less time and 18% higher quality on writing tasks with consumer ChatGPT among 453 professionals.

The Dillon and colleagues Copilot RCT (NBER w33795, November 2025 revision) is the most important causal evidence on Microsoft 365 Copilot as an enterprise product. Among 7,137 knowledge workers in 66 firms, treated workers who used the tool (80% of treated in the second half of the six-month experiment) spent two fewer hours on email each week in the local-average-treatment-effect sense emphasized in the abstract; the intent-to-treat Outlook reduction is 1.4 hours per week (12%). The paper does not detect shifts in the quantity or composition of workers' tasks from individual-level AI provision. Meeting time did not significantly change in the main stated results. Several authors were Microsoft employees. This is simultaneously evidence of time savings and of a null effect on measured task output—an essential mixed finding.

Dell'Acqua and colleagues' BCG experiment shows +12.2% more tasks completed and 25.1% greater speed inside the frontier, and a 19 percentage-point lower likelihood of correct solutions on an outside-frontier task. Their P&G experiment ($n = 791$) finds that individuals with AI matched the performance of teams without AI on the paper's stated measures. Otis and colleagues (Kenya, $n = 640$) cannot reject a zero average effect on entrepreneurial profits and find worse relative outcomes for low baseline performers—an international caution against assuming positive average returns in open-ended work. Ke and colleagues (Singapore, 2025) find no significant overall documentation-time reduction from a perioperative chatbot, with subgroup minutes saved. Afshar et al. [50] find a 0.36 hour per day reduction in notes time (95% CI -0.55 to -0.17) with an ambient AI scribe.

Large public-sector Copilot evaluations are self-report dominant. The Government Digital Service evaluation [55] (20,000 licenses; 7,115 survey responses) reports a mean 26 minutes saved per day, with 17% reporting no clear savings and more than one-third saving more than 30 minutes; official annualization is 13 days per year. The government news release's

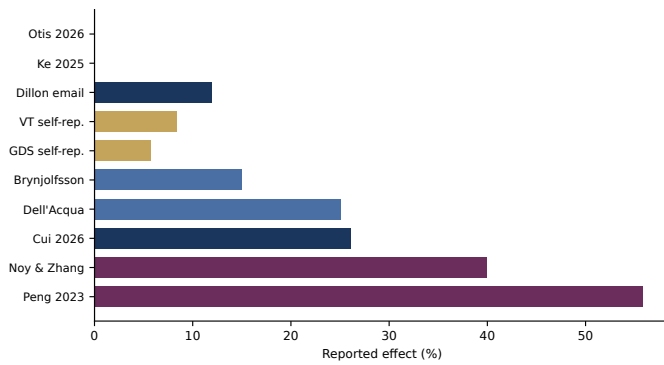


Fig. 6. Structured comparison of selected productivity and time-saving estimates. This is not a meta-analysis and does not report a pooled effect, confidence interval, or I^2 . UK GDS and Virginia Tech minute figures are illustrated as a share of a 7.5-hour day solely for visual scale; that conversion is a reviewer annotation, not a source statistic. Virginia Tech is vendor-partner and is excluded from the independent UX synthesis.

TABLE X

SELECTED PRODUCTIVITY AND TIME-SAVING OUTCOMES. ESTIMATES ARE NOT POOLED. COMPLETER-ONLY AND SELF-REPORT ITEMS ARE LABELED.

Study	Outcome	Estimate	Design
Brynjolfsson et al. [16]	Issues/hr	+15%	Field
Cui et al. [18]	Tasks	+26.08% (SE 10.3%)	3-firm RCT
Peng et al. [64]	Time	-55.8% (completers)	Lab RCT
Noy and Zhang [17]	Time/quality	-40% / +18%	Online RCT
Dell'Acqua BCG	Speed/accuracy	25.1% / -19 pp outside frontier	RCT
Dillon et al. [49]	Outlook/tasks	-12% / null task shift	RCT
Afshar et al. [50]	Notes time	-0.36 h/day	RCT
[55]	Self-report	26 min/day	Field
Virginia Tech	Self-report	38 min/day (n=85)	Pilot
Ke et al. [65]	Doc. time	Null overall	RCT

“nearly two weeks” language is that annualization of the same self-report, averaged in the news text across the 20,000-license cohort. There was no control group, and researchers could not identify how saved time was spent. A U.S. university volunteer pilot reports 38 minutes average daily savings among 85 closing-survey respondents, with 94% of those respondents reporting some daily saving (Cloudforce/Microsoft-partner pilot; excluded from independent UX synthesis). CSIRO’s 300-license trial shows more modest Likert agreement on productivity (21% agree, 37% somewhat agree). These pilots are useful and not causal in the Dillon sense. Fig. 6 and Table X compare the reported productivity and time-saving effects across studies without pooling them into a single estimate.

4.6 RQ5 — ROI and business value

RQ5 asks whether enterprise AI is actually making money for organizations, or whether most reported returns are only estimates, models, or self-reported claims. We did not identify an included study that independently measured a financial ROI for a named enterprise assistant or agent deployment using audited cost and benefit ledgers. McKinsey [52] finds that 39% of respondents report any enterprise EBIT impact from AI and that most of those impacts are under 5% of EBIT; about 6%

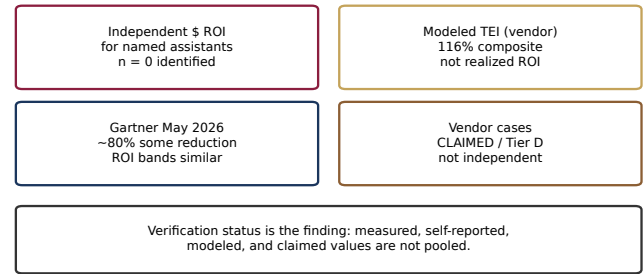


Fig. 7. ROI-like claims classified by verification status. Independently measured financial ROI of named enterprise assistants: $n = 0$.

TABLE XI

ROI AND BUSINESS-VALUE CLAIMS. INDEPENDENTLY VERIFIED FINANCIAL ROI OF NAMED ENTERPRISE ASSISTANTS: $N = 0$.

Source	ROI claim	Verification	Tier
Independent studies	None identified	—	—
[52]	39% any EBIT impact; most <5%	Self-report	C
Deloitte	20% claim ROI >30%	Self-report	C
[24]	Cuts not linked to higher ROI	Self-report	C
Forrester TEI [28]	116%; \$19.7M NPV	Commissioned model	D
NANDA 2025	95% zero return	Preliminary	C
Klarna PR	\$40M estimate	Vendor claim	D

are “high performers” on McKinsey’s definition. Deloitte Wave 4, among organizations already piloting generative AI, reports that almost all claim measurable ROI on their most advanced initiative, that 20% claim ROI above 30%, and that 74% meet or exceed expectations—self-report in a selected sample. The MIT NANDA “GenAI Divide” report (July 2025) states that 95% of organizations are getting zero return despite \$30–40 billion of enterprise investment, based on 300+ public initiatives, 52 interviews, and 153 conference-surveyed leaders. The authors themselves note that percentages reflect the interview sample of 52 organizations and label findings preliminary. We include the 95% figure only with those constraints and do not treat it as a population parameter.

Gartner’s May 2026 result [24] that workforce reductions do not track ROI is the most important negative finding on the substitution-for-value hypothesis in the analyst tier. The Forrester TEI 116% ROI [28] is a composite. Vendor case studies (OpenAI, Anthropic, ServiceNow, Salesforce/Futurum) report selected dollar and FTE-equivalent claims and are coded D/CLAIMED. Fig. 7 and Table XI summarize verification status.

4.7 RQ6 — User satisfaction and adoption intensity (independent synthesis)

This subsection reports only INDEPENDENT evidence on real users. Populations are not combined.

4.7.1 A. Employees using an AI assistant (Microsoft 365 Copilot): Five independent public-sector evaluations surveyed or compared employees who actually used licensed Microsoft 365 Copilot. Findings are mixed by instrument but lean positive

on satisfaction and continued-use intention among people who received licences and responded.

UK GDS (S027; 20,000 licences; survey $n=7,115$; telemetry $\sim 14,500$): 82% would not want to return to pre-Copilot work; satisfaction 7.7/10; recommendation 8.2/10; 85% agreed the tool had good organizational value. Active use peaked at 83% and remained about 80% (30-day definition). Feature use was uneven (Teams peak 71% active; Excel 23%; PowerPoint 24%). Seventeen percent reported no clear time savings. Microsoft supplied usage-dashboard telemetry; GDS authored the analysis.

DWP (S049; 3,549 licences; 1,716 users and 2,535 non-users): 95% of users were very or somewhat satisfied; 65% reported daily use and 30% weekly; 77% were extremely interested in licence extension (versus 50% extremely interested when first notified). SUR estimates +0.56 points on a 7-point job-satisfaction scale versus non-users after controls. Ninety percent said the tool saved time; SUR time saving was 19 minutes per day. Interviews described the tool as easy to use for many Microsoft-familiar staff; prompting and accuracy still required human edit. Some users raised job-security and dependency concerns. This is not a randomized licence trial.

DBT (S050; 1,000 licences; diary $\sim 32\%$ response; NPS calculated on diary respondents): 72% satisfied or very satisfied; NPS 31; overall dissatisfaction under 1%. Slide-deck work was 54% satisfied (23% dissatisfied) and scheduling 57% satisfied (36% dissatisfied). The evaluation did not find evidence that time savings translated into improved productivity; control interviewees had not observed productivity improvements in colleagues. Usage averaged 1.14 Copilot actions per user per working day (Microsoft dashboard interpreted by DBT).

HMRC Phase III (S051; 3,000 randomly allocated licences plus 500 additional; survey 1,364): 83% of allocated staff used the licence; mean satisfaction 7.1/10; 68% agreed it was easy to use and integrated; 61% would be disappointed to lose the licence; 64% said use increased during the trial. Among non-users, 46% cited security and data-privacy concerns as the main reason. Classification rules barred Official-Sensitive material.

CSIRO / Bano and colleagues (S030; 300 licences; post-survey 171): recommendation 7.13/10 (SD 2.44). Productivity agreement was mixed (21% agree / 37% somewhat agree more productive). Privacy concerns rose after the trial. A Microsoft observer sat on a related Digital Transformation Agency program board after probity advice; authors are CSIRO researchers and the surveys are treated as independent with that disclosure.

What users liked most, where measured: embedding in familiar Office apps; drafting, summarising, search, and meeting capture; accessibility benefits for some neurodivergent staff and non-native English speakers (DWP/DBT qualitative). Frequent complaints: uneven quality on slides and scheduling (DBT); Excel and some apps unused (GDS); prompting/learning curve; security/privacy non-use (HMRC); weaker performance on nuance-heavy work (GDS); no demonstrated productivity translation (DBT). Abandonment in these samples is more often unused features or unused licences than wholesale walk-away among survey respondents. Samples are licensed users, not all

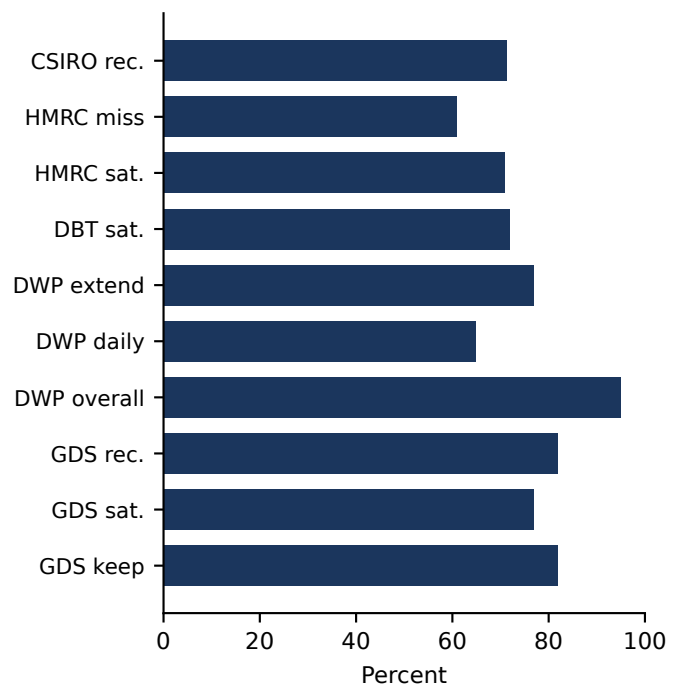


Fig. 8. Independent licensed-user Copilot satisfaction and continued-use items (UK GDS, DWP, DBT, HMRC; CSIRO). Vendor-affiliated pilots omitted. 0–10 scores are shown as score $\times 10$ for scale only.

employees.

4.7.2 B–E. Configurers, citizen developers, professional developers, IT/security administrators: Insufficient independent empirical evidence was identified on satisfaction or usability among business users configuring Copilot, citizen developers building agents in Copilot Studio or Power Platform, or professional developers integrating those agents. Independent institutional sources document governance and security friction for administrators (GAO policy drag; HMRC classification and privacy-driven non-use; GDS permission inheritance) but do not survey administrator satisfaction. Insufficient comparable independent evidence exists for ChatGPT Enterprise, Gemini, Agentforce, ServiceNow, Claude Enterprise, or AWS agent-builder user experience after vendor-affiliated reports are removed. No ranking across platforms is offered.

4.7.3 Other independent assistant users (not Copilot): Afshar et al. (S009) [50] measured clinician well-being with an ambient scribe (Abridge), not Copilot; vendor reviewed the manuscript but authors disclosed no AI-vendor research funding. That trial is independent well-being evidence for a different product and is not pooled with Copilot satisfaction percentages. Brynjolfsson and colleagues (S002) and Dell’Acqua and colleagues (S006) measure task performance, not satisfaction instruments, and are not treated as UX surveys. Fig. 8 and Table XII summarize independent Copilot employee items only. Fig. 9 records qualitative friction codes; no numerical friction scores were invented.

TABLE XII
INDEPENDENT USER-EXPERIENCE RESULTS USED IN THE PRIMARY SYNTHESIS. POPULATIONS ARE NOT COMBINED. BUILDER ROWS REPORT AN EVIDENCE GAP, NOT A ZERO SCORE.

Setting	User type	Item	Value	Independence
UK GDS	Employee	Would not return to pre-Copilot work	82%	INDEPENDENT
UK GDS	Employee	Satisfaction / recommend	7.7 / 8.2 of 10	INDEPENDENT
UK GDS	Employee	30-day active use	Peak 83%; ~80% maintained	INDEPENDENT
DWP	Employee	Satisfied with performance	95%	INDEPENDENT
DWP	Employee	Daily use / want extension	65% / 77% extremely interested	INDEPENDENT
DBT	Employee	Satisfied / NPS	72% / 31	INDEPENDENT
HMRC	Employee	Satisfaction / miss licence	7.1/10 / 61%	INDEPENDENT
CSIRO	Employee	Recommend	7.13/10 (SD 2.44)	INDEPENDENT
Copilot Studio builders	Citizen / pro developer	Satisfaction or complexity RCT/survey	Insufficient independent empirical evidence was identified.	—

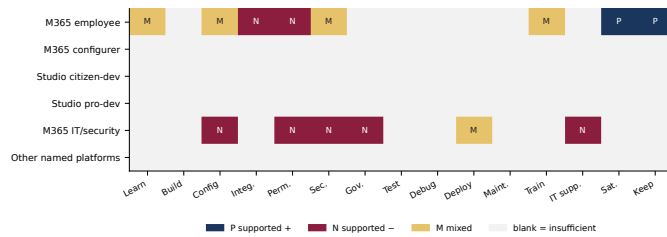


Fig. 9. User friction matrix for enterprise assistants and agent-builder stages. Cells are qualitative codes (supported / mixed / insufficient evidence) from independent sources only. No numerical friction scores were assigned. Comparable independent evidence across named vendor platforms remains insufficient.

4.8 RQ7 — Training, implementation difficulty, and building complexity

Independent employee studies describe training as uneven rather than as a measured hour–outcome function. GDS: departmental support inconsistent; familiarity correlated with reported time savings. DWP: 89% used self-directed exploration; 77% used departmental resources; users wanted shorter, role-specific sessions and cheat-sheets. DBT: self-led training was more strongly associated with satisfaction than departmental sessions. HMRC: practical, tailored training and time to learn were explicit challenges. CSIRO does not report a dose–response. The Forrester TEI study [28] modeled \$6.9 million training cost is vendor-commissioned and is not used here.

Building complexity by stage (independent evidence only): initial learning MIXED for employees using Copilot; prompt design MIXED; connector/API integration SUPPORTED NEGATIVE where Excel and some apps are unused or poor (GDS/DBT); permissions and security SUPPORTED NEGATIVE (ACL inheritance; Official-Sensitive limits; privacy as a non-use reason). Agent creation, RAG, orchestration, authentication, testing, debugging, monitoring, and maintenance: Insufficient independent empirical evidence was identified. No-code claim validation: independent sources describe LEVEL 2 departmental assistant use. They do not show that nontechnical employees independently progress to LEVEL 3 integrated production agents or LEVEL 4 governed enterprise agents in Copilot Studio. Whether professional IT/security intervention

is required at those later levels remains an evidence gap, not a vendor-claim substitute.

4.9 RQ8 — Reliability and performance limitations

RQ8 examines the reliability of enterprise AI and the conditions under which its performance becomes limited or less accurate. Reliability problems are documented in both laboratory and field settings. Dell’Acqua and colleagues provide the cleanest causal demonstration that AI can reduce accuracy outside the frontier. Dillon and colleagues show time savings without detected task-volume change, which is a performance limitation relative to transformation claims. UK GDS reports weaker performance on nuance- and context-heavy work and difficulty of Copilot agents (mid-trial, formally out of scope) in identifying exact source documents. Virginia Tech reported that Copilot agents released in March 2025 were not sufficiently tested and that Excel advanced analytics were unavailable in-tenant. Debenedetti et al. [41] find benign-task utility below 66% even without attack. Ke and colleagues’ null average clinical documentation effect is a reminder that “AI assistant” does not imply time savings. Hallucination rates in production Microsoft 365 Copilot or ChatGPT Enterprise tenancies are not independently quantified in the included organizational studies.

4.10 RQ9 — Cybersecurity, privacy, and governance

NIST AI RMF 1.0 [29] (Govern–Map–Measure–Manage) and the Generative AI Profile [30] (July 2024) are voluntary U.S. frameworks, not outcome evaluations. NIST AI 100-2e2025 [31] defines prompt injection and indirect prompt injection. AgentDojo [41] and the NIST/CAISI 17 January 2025 technical blog [32] provide empirical attack results: CAISI reports a jump from an 11% strongest-baseline attack success rate to 81% with a stronger attack on a Workspace hold-out, a five-task average of 57%, and 80% when attacks were repeated 25 times on Claude 3.5 Sonnet (October 2024). These are laboratory rates, not counts of enterprise incidents.

In deployments, UK GDS states that Copilot inherits user permissions, that internet search was off by default in most participating organizations, and that security or sensitive-data concerns reduced benefits for a minority. Virginia Tech



Fig. 10. Security and governance risks with empirical laboratory support or official deployment documentation. Laboratory attack-success rates are not organizational breach rates. Sources: Debenedetti et al. (2024); NIST CAISI (17 January 2025); UK GDS (2025); GAO-25-107653; OMB M-25-21.

TABLE XIII
SECURITY AND RELIABILITY CHALLENGES DISTINGUISHED BY WHAT IS KNOWN AND WHAT REMAINS UNKNOWN. LABORATORY ATTACK-SUCCESS RATES ARE NOT INCIDENT COUNTS.

Risk class	What is known	What is unknown	Sources
Prompt / indirect injection	Defined (NIST); high lab ASR possible	US enterprise incident rate	S022, S010, S023
Excessive permission / ACL inheritance	Documented product behavior	Frequency of oversharing events	S027, S028
Shadow AI / BYOAI	Vendor survey only (excluded from independent UX)	Independent incidence unknown	S045 (vendor-affiliated)
Privacy policy friction	10/12 GAO agencies	Quantified delayed value	S018
DLP / logging / residency	Required in policy talk	Comparative product audits	S020, S024

emphasizes tenant containment suitable, in its account, for high-risk data. GAO [37] reports that officials at 10 of 12 selected agencies said existing federal policy, including privacy policy, could impede adoption. OMB M-25-21 [38] directs agencies to develop a generative AI acceptable-use policy within 270 days. A Microsoft/LinkedIn Work Trend Index BYOAI statistic is vendor-survey evidence and is not used in the independent UX or security incidence synthesis. We did not find included independent measurements of production data-leakage incidence for Copilot, ChatGPT Enterprise, or Agentforce in U.S. organizations. Fig. 10 and Table XIII separate laboratory attack results from organizational unknowns.

4.11 RQ10 — Human oversight and backup

Human-in-the-loop is the modal design in the strongest studies. Customer-service agents could ignore suggestions (35% follow rate). Clinicians remained accountable for notes. Copilot users remained authors of email and documents. UK GDS states that human oversight was required at all times. Klarna's 2025 recalibration is the clearest public example of restoring human access after an automation-first period produced, in the CEO's description, lower quality. We found little detailed public evidence on rollback procedures, compensation for incorrect autonomous actions, or business-continuity runbooks specific to agentic misfires in U.S. enterprises. Table XIV records what is present and absent.

4.12 RQ11–RQ12 — Challenges, success, and failure

Recurrent challenges with multi-source support include: inability to demonstrate independently verified financial ROI; pilot-to-production stall (Deloitte: more than two-thirds say 30% or fewer experiments fully scale in three to six months; [66] prediction that over 40% of agentic projects will be canceled by the end of 2027 because of cost, unclear value, or inadequate risk controls—a prediction, not an observed rate); self-reported minutes that do not identify reallocated capacity (UK GDS); security and privacy policy drag (GAO); uneven training; jagged quality; and public-sector AI talent shortages (GAO), including the IRS skills-gap case. Vendor lock-in is discussed in practitioner sources but is not measured as a switching-cost study here.

Success, defined as independently measured improvement on a pre-specified operational metric without a corresponding quality collapse, is observed in several Tier A settings (customer-service IPH, developer tasks, some writing and consulting tasks, ambient documentation time). Mixed success includes Copilot's email-time reduction with null task-composition change, BCG's inside/outside frontier split, and high satisfaction with unknown time-use in government. Failure or reversal includes Klarna's quality-driven human return, NANDA's account of custom tools stalling at 5% production, Ke and colleagues' null average time effect, Otis and colleagues' null average profit effect, and Gartner's forecast of widespread agentic cancellation. We do not have a large enough set of identically measured deployments to estimate a success rate. Fig. 11 codes recurrent challenges by breadth of independent source classes, not as prevalence

TABLE XIV
HUMAN OVERSIGHT AND FALLBACK MECHANISMS DOCUMENTED IN INCLUDED SOURCES. UNKNOWN CELLS ARE EVIDENCE GAPS, NOT PROOF THAT NO MECHANISM EXISTS.

Setting	Oversight model	Fallback / backup	Evidence
Brynjolfsson CS	Human-in-the-loop suggestions	Work continues if ignored	DIRECT
Afshar ambient scribe	Clinician remains legally responsible	Usual documentation in control periods	DIRECT
UK GDS Copilot	Oversight required at all times	Ordinary civil-service process	DIRECT (stated)
Dillon Copilot RCT	User retains application control	Control arm without Copilot	DIRECT
Klarna CS agent	Initially automation-heavy	2025 human access restored (news)	CLAIMED/ INDIRECT
Production Agentforce / Copilot Studio autonomous actions	Not independently documented in included set	Not reported	UNKNOWN

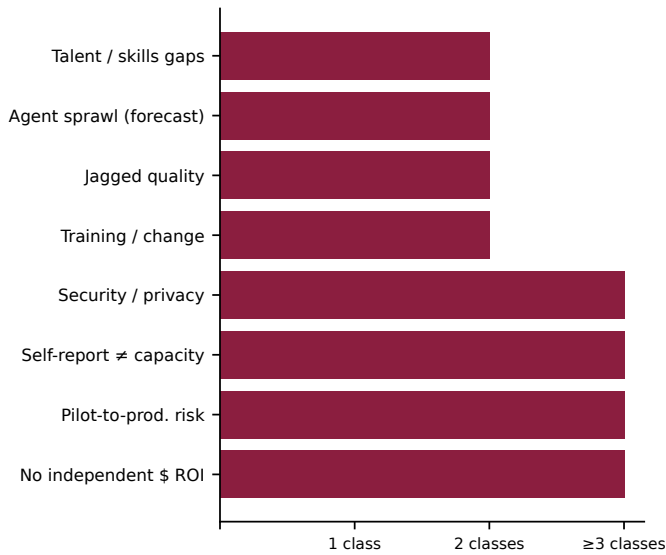


Fig. 11. Recurrent implementation challenges coded by the breadth of independent source classes that support them, not as prevalence rates in a defined population.

rates.

4.13 RQ13 — AI investment versus workforce change

The organization-level dataset contains 18 identifiable organization or program rows. None contains both a documented dollar AI expenditure and an independently verified subsequent layoff wave attributed to that expenditure. The IRS row has workforce reduction with causality code 4. Klarna has claimed FTE-equivalents and later human reinvestment. Census evidence is a cross-section of firm shares, not a panel linking invoices to headcount. Fig. 13 therefore reports a data-availability result: $n = 0$ complete pairs. We do not compute a correlation. Computing one would violate the protocol’s prohibition on inferring causality—or even association—from missing data. Table XV, Fig. 12, and Section 4.13.1 extend that finding along the 2022–2026 chronology and the A–F attribution ladder.

Budget displacement—staffing reduced because money moved to AI rather than because AI automated the work—is a proposed analytical category (Section 4.15.1), not an official CGEIT/GAO class. What the longitudinal record actually shows: Microsoft’s \$64.551 billion FY2025 PPE additions are vendor infrastructure outlays, not buyer salary displacement. GSA promotional prices and a \$3.1 billion claimed savings figure are not paired with federal FTE cuts in those releases. IBM’s May 2023 hiring pause is a stated intention to let AI and automation absorb back-office work, including via attrition, before Copilot GA—AI-contributing planning, not a documented 7,800-person 2023 layoff and not an invoice-to-headcount ledger. The IRS case is the clearest public example of staff reductions coinciding with an AI program, and GAO attributes the 63 losses to broader 2025 staffing cuts, warning that losing AI specialists raises the risk that AI efforts will fail—the opposite of a substitution dividend. Gartner’s finding that reduction rates were similar across ROI bands is consistent with cuts that are not benefit-justified, but it does not prove mechanism (3) in any named firm. Where AI investment and workforce reductions occur in the same fiscal window without a documented trade-off, the correct label remains temporal association only; causal attribution not established. That chronology is why Section 4.15.1 requires a documented mechanism before a board may label a budget or headcount action “AI-caused.”

4.13.1 Longitudinal AI Adoption–Spending–Workforce Attribution Framework: Chronology is not causality. This subsection asks, for each major U.S.-relevant case, which links in the chain are evidenced and which are missing: AI announcement → enterprise availability → organizational procurement/spending → actual deployment/use → measured productivity or automation → job-posting or hiring change → workforce reduction or layoff → management explanation. Classes are A (direct AI evidence of task or role elimination), B (AI one documented factor), C (investment-induced budget displacement—proposed category), D (general restructuring), E (temporal association only), and F (unknown/insufficient). No case in this search is labeled AI-caused (class A) at the level of a completed, independently verified named-role substitution tied to a named enterprise agent invoice.

TABLE XV
AI SPENDING VERSUS WORKFORCE-CHANGE EVIDENCE. TYPES OF MONEY (ACTUAL, COMMITTED, MODELED, LIST, CLAIMED) ARE NOT COMBINED.

Pair	What is observed	What is not established
MSFT 10-K PPE vs customer TCO	Actual vendor capex \$64.551B FY25	Buyer Copilot invoices or buyer layoffs
GSA OneGov vs federal FTE	Promo prices; \$3.1B claimed	Employment effect
Census use vs 2% decrease	Same survey wave	Invoices or worker-level causation
IBM pause vs AI	CEO attributes pause to AI/automation	Completed 7,800 layoffs or license TCO
IRS 63 vs AI program	Staff lost; AI plans continue	Direct substitution (GAO: code 4)
Gartner cuts vs ROI	Org-share reductions; flat ROI bands	Named-firm budget displacement
Complete invoice + layoff	n = 0	Any correlation

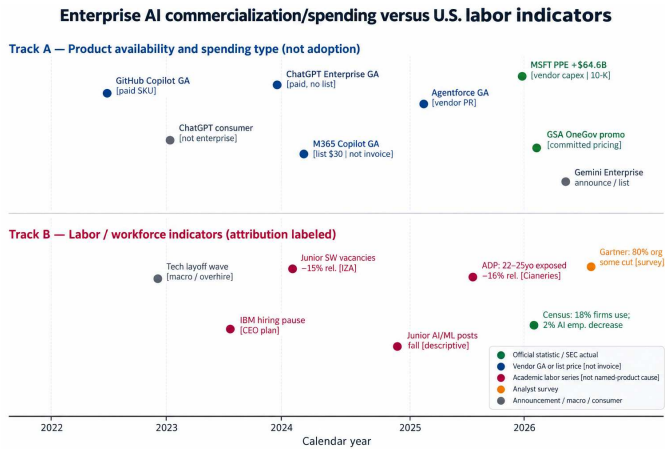


Fig. 12. Dual-track chronology, 2022–16 September 2026. Track A: product availability and spending type. Track B: labor indicators. Colors mark official/SEC actual, vendor GA or list price, academic labor series, analyst survey, and announcement/macro events. No causal line is drawn from A to B. Data sources are listed in the companion longitudinal workbook.

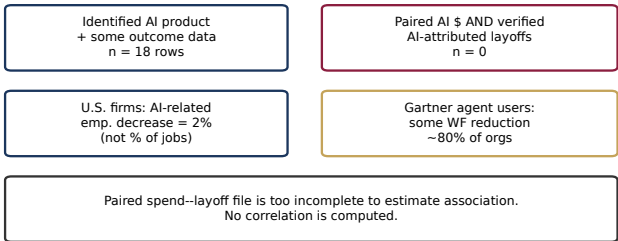


Fig. 13. Availability of paired evidence on documented AI expenditure and independently verified subsequent workforce reduction. Complete pairs in the organization-level file: n = 0. No correlation or causal model is estimated. Sources: Organization_AI_Workforce_Dataset.xlsx; CES-WP-26-25; Gartner (5 May 2026).

What the chain actually shows through 16 September 2026. Paid enterprise assistants become purchasable in 2023 (ChatGPT Enterprise; Microsoft 365 Copilot GA). Vendor agent platforms become purchasable in 2024–25. National firm use remains a minority in early 2026. The only national official employment item is a 2% firm-share self-report of AI-related decrease (class F at worker level). IBM’s May 2023 pause is class B planning before Copilot GA, with SKU, invoice, measured automation, and completed layoff count missing. The IRS 63-person loss is class D: GAO attributes it to broader 2025 staffing cuts and treats lost AI skill as a threat to AI success. The Dillon Copilot RCT and the QJE customer-service

assistant study fill deployment and productivity links and leave the layoff link empty. Gartner organization-share items are class B surveys, not employee-share causation and not class C in any named firm. 2022–23 technology layoffs are class E/D: they largely precede Copilot GA and are explained in filings and press as rates, overhiring, and growth reset. Microsoft’s \$64.551 billion FY2025 PPE additions are vendor capex, not buyer displacement. Where AI spend and headcount change share a fiscal window without a documented trade-off, the required label is: Temporal association only; causal attribution not established. Section 4.4.2 adds the indirect-mechanism ladder (classes 1–7). IBM and the 2025 Salesforce engineer-hiring statement are class 4 (and class 2 language), not class 1 completed layoffs. Duolingo contractor non-renewal is class 5. Bank of America attrition mix is class 5. Class 3 explicit budget displacement remains n = 0 as a complete ledger. Indirect compression appears more often than explicit AI-layoff counts in the named-org record; the large 2022–23 cuts remain class 6/7.

Job postings. JOLTS nonfarm openings fell from a March 2022 peak near 12.0 million toward later, lower levels (BLS; the exact end month depends on the cited JOLTS vintage table). That decline is a macroeconomic / post-pandemic normalization series, not an AI attribution. Lightcast software-developer vacancies (SOC 15-1252) show a roughly 14–15% relative junior decline around ChatGPT and, descriptively, junior AI/ML postings 1,915 → 1,029 (-46.3%) versus senior 4,338 → 3,934 (-9.3%) from October 2022 to October 2024 [61]. ADP payrolls show a 16% relative employment decline for ages 22–25 in the most AI-exposed occupations through September 2025 after firm-time controls [62]. Confounders that remain on the table: interest rates, tech overhiring correction, outsourcing, recession expectations, and industry contraction. Comparable monthly posting extracts for customer service, administrative support, HR/recruiting, marketing/content, finance/accounting, and data/analytics were not compiled as a complete panel in this review and are not invented. A posting decline is not, by itself, AI-caused. Table XVI records the case-level attribution chain.

4.14 RQ14 — Copilot-specific synthesis

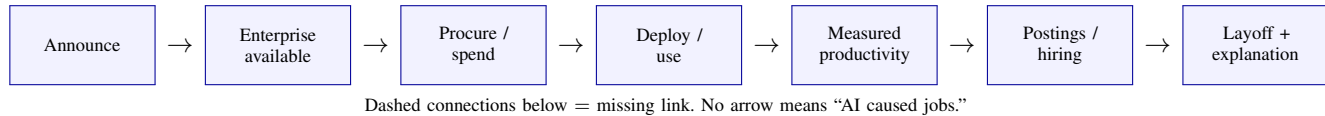
Microsoft 365 Copilot remains the only named enterprise assistant with both a multi-firm RCT (productivity/time-use; Microsoft-affiliated authors—usable for causal time-use, not for independent satisfaction) and large independent public-sector user evaluations. Cost (not a UX claim): commercial list

TABLE XVI

LONGITUDINAL ATTRIBUTION. FULL COLUMNS—including SPENDING, PRODUCTIVITY, POSTINGS, MANAGEMENT WORDING, ALTERNATIVES, EVIDENCE STRENGTH, AND EXACT SOURCE—are IN ENTERPRISE_AI_LONGITUDINAL_EVIDENCE.XLSX. CLASS A WAS NOT ASSIGNED.

Org / sector	Tech	Class	Links present	Links missing
Census firms	Any AI	F (worker)	Use; 2% decrease item	SKU; invoice; verified layoff
IBM	Unspecified AI/automation	B planned	CEO hiring-pause explanation	SKU; TCO; completed 7,800
IRS RAAS	IRS AI program	D	63 staff lost; GAO explanation	Substitution; AI invoice
Copilot RCT firms	M365 Copilot	F workforce	Deploy; -12% email time	Layoff; invoices
F500 CS (QJE)	GPT reply assistant	F workforce	Production +15% IPH	Layoff; agentic writes
Gartner n=350	Agents (self-report)	B	Org-share cuts; flat ROI	Invoices; class C proof
IZA / ADP	Exposure / ChatGPT timing	B (E if product-causal)	Junior demand / 22–25 emp.	Named agent; invoices
2022–23 tech wave	Often none	E / D	Cuts; macro explanation	Copilot invoices
Invoice + layoff pair	—	F	None	Entire paired ledger (n=0)

Evidence chain (links present $\neq$ causation)



- **IBM 2023 pause [B]** — Explain + hiring intent; missing SKU, invoice, and completed layoff evidence.
- **IRS RAAS 63 staff [D]** — Cut + audit; missing substitution mechanism and AI invoice.
- **Dillon Copilot RCT [F/wrk]** — Deploy + time use; missing layoff and invoices.
- **QJE CS assistant [F/wrk]** — Production assist +15% IPH; missing layoff.
- **Gartner n=350 [B]** — Self-report use + org-share cuts; missing invoices.
- **IZA / ADP series [B/E]** — Labor demand/employment; missing named agent.
- **Invoice + layoff pair [F]** — $n = 0$ complete ledgers.

Classes: A direct; B contributing; C budget displacement (proposed); D general cuts; E temporal only; F unknown. Announcement → enterprise availability → procurement/spending → deployment/use → measured productivity/automation → labor-posting change → workforce reduction → management explanation.

Fig. 14. Which evidence-chain links are present. Colors mark attribution class. No arrow is a causal claim that AI destroyed jobs. Companion to Fig. 12 (synchronized tracks; official / company / research / survey coded; no causal A→B line).

\$30 per user per month plus a qualifying suite; U.S. federal G5 promotional zero-price year. Independent employee UX: high satisfaction and continued-use interest in GDS, DWP, DBT, HMRC, and CSIRO samples, with mixed task usability (slides, scheduling, Excel) and mixed translation of minutes into productivity (DBT: no evidence of productivity improvement). Dillon and colleagues show email-time reductions and no detected task-quantity shift; that study is vendor-affiliated for UX purposes and is cited here only as productivity telemetry. Workforce: no included independent Copilot evaluation reports layoffs caused by Copilot. Security/governance: permission inheritance, classification limits, and privacy-driven non-use are independently documented. Copilot Studio / Power Platform agent building: Insufficient independent empirical evidence was identified. GitHub Copilot task RCTs have Microsoft-affiliated authors and are not used as independent developer-UX evidence. ChatGPT Enterprise telemetry (S031) is vendor-affiliated. Cross-platform ranking: Insufficient comparable independent evidence.

4.15 RQ15 — CGEIT-aligned governance synthesis

Fig. 15 and Table XVII (CGEIT matrix) organize evidence without scoring. Across deployments, technological access is often easier to observe than benefits realization or residual-

risk optimization. Federal use-case counts and OneGov price cuts show rapid availability; GAO simultaneously documents policy friction and talent gaps. Census adoption (18% of firms) coexists with only 2% reporting AI-related employment decreases and with no independent ROI ledger. Gartner’s large-firm agent users report cuts that do not buy higher ROI. Independent UK Copilot evaluations show relatively stronger governance and oversight language and relatively stronger (if still self-reported) employee satisfaction, without financial verification. Klarna shows claimed benefits followed by risk and quality problems when human fallback was insufficient. The IRS shows resource destruction (skills) that GAO links to higher risk of unsuccessful AI, not to successful substitution.

Is adoption occurring faster than organizations’ ability to govern, demonstrate benefits, optimize risk, and manage IT resources? Several included sources are consistent with that descriptive claim—McKinsey’s gap between 88% functional use and 39% any EBIT impact; Deloitte’s scale-up failure share; NANDA’s production stall; Gartner’s predicted agentic cancellations; GAO’s policy and skills lags; Dillon’s null task-output result despite license provision. The claim is not proven as a causal law, and some deployments (structured CS assistance, ambient scribes, developer completions) do demonstrate measured operational benefits with human oversight. Fig. 16

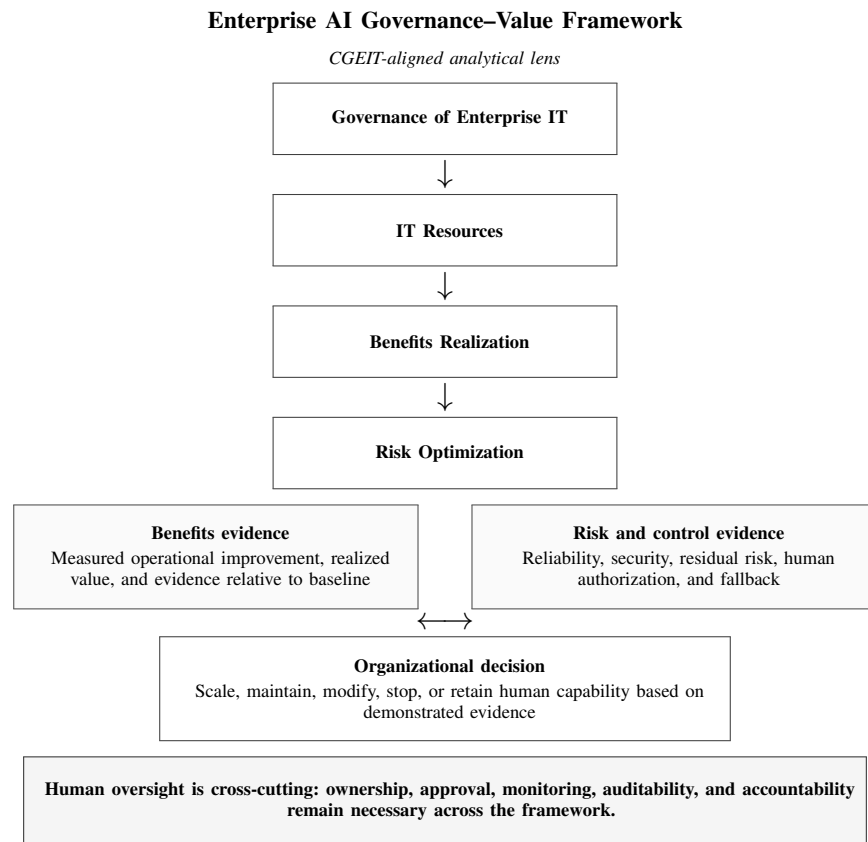


Fig. 15. Enterprise AI governance–value framework aligned to the four CGEIT knowledge domains, with human oversight as a cross-cutting requirement. The figure is an interpretive scaffold for Section 4.15, not a maturity score or certification model.

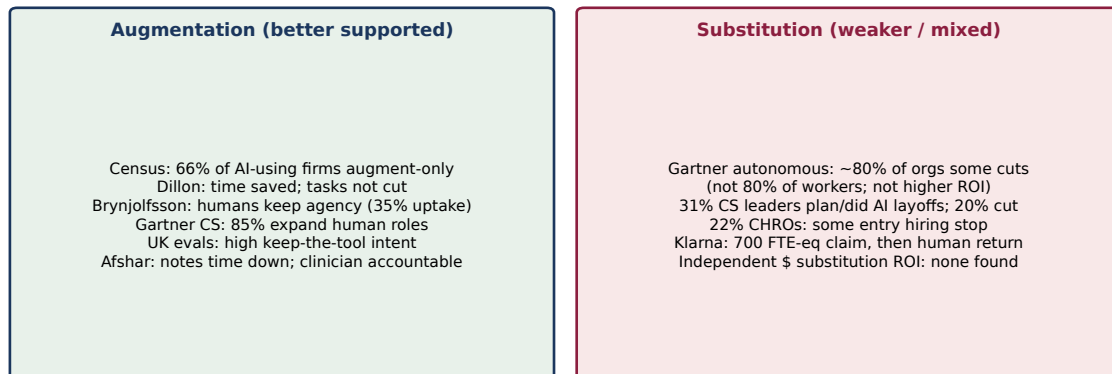


Fig. 16. Qualitative mapping of included evidence toward augmentation versus substitution. The figure is conceptual and is not a fitted statistical model.

maps that qualitative pattern. Successful cases in this set more often have a narrow task, a human decision right, and a measurable operational metric. Unsuccessful or mixed cases more often have open-ended goals, headcount-first value stories, or absent measurement. That differentiation is supported as a qualitative synthesis, not as a regression.

4.15.1 Evidence-Gated Organizational Decision-Making for Enterprise AI: RQ15 asks whether recognized governance frameworks provide usable guidance for major organizational decisions after enterprise AI assistants or agents are deployed—budget allocation or expansion, enterprise-wide license expansion or renewal, pilot-to-production transitions, workforce restructuring, hiring freezes, role elimination, layoffs attributed

to AI, and decisions based on claimed productivity, ROI, cost savings, or automation benefits. This subsection reports what official sources actually require or recommend, then labels a decision-stage evidence model as a proposal of this review. CGEIT remains an interpretive lens; no organization is described as CGEIT-compliant. The 2022–2026 chronology in Tables III and XVI strengthens the gate: paid enterprise SKUs became available in 2023, vendor and federal money is visible mainly as infrastructure capex and promotional prices, junior labor-market series deteriorated around the same years for reasons that include but are not limited to AI exposure, and still no independently verified invoice-to-layoff pair exists. That pattern is why claimed minutes, list prices, or a commissioned

TABLE XVII

CGEIT-ALIGNED ENTERPRISE AI GOVERNANCE EVIDENCE MATRIX (ABBREVIATED). LABELS INDICATE WHETHER INCLUDED SOURCES SUPPORT A DOMAIN DESCRIPTIVELY; THEY ARE NOT CERTIFICATION SCORES. VIRGINIA TECH IS RETAINED AS A GOVERNANCE CASE AND IS EXCLUDED FROM THE INDEPENDENT UX SYNTHESIS. FULL TEXT FIELDS ARE IN QUALITY_ASSESSMENT.XLSX.

Organization	Gov.	Resources	Benefits	Risk	Oversight	ROI	Workforce	Tier
VT Copilot	Supported	Partly (no TCO)	Measured survey	Partly	Supported	Not reported	None (5)	B
UK GDS	Supported	Partly (uneven training)	Measured survey	Partly	Supported	Not \$	None (5)	B
Dillon RCT firms	Not reported	300-license incentive	Time yes; tasks null	Not primary	User control	Not reported	Not measured	A
F500 CS firm	Not fully	Tool + 35% uptake	IPH +15%	Quality mix	HITL	Not reported	No layoff finding	A
IRS	Weak (GAO)	Lost 63 AI staff	Not shown	Gaps	Fragmented	Not reported	Cuts code 4	B
Klarna	Mixed	Automation then rehire	Claimed \$40M	Quality risk	Restored later	Claimed	Code 2	D
US firms (Census)	Not measured	18% adopt	Not ROI	Not measured	Augment modal	Not reported	2% of firms	B
Gartner autonomous	Not assessed	Spend forecasts	ROI≠cuts	Not primary	Recommended	Self-report	80% of orgs	C

ROI model are not sufficient evidence for enterprise-wide scale-up or irreversible workforce actions.

What official frameworks require or recommend

ISACA's CGEIT exam content outline (effective 2020) is a job-practice specification for governance professionals, not a statute and not an AI-product regulation. Domain 2 (IT Resources, 15%) covers IT resource planning and optimization, including resource capacity planning, acquisition, the IT-resource lifecycle, and human-resource competency assessment (ISACA, CGEIT Examination Content Outline, Domains 2A–2B). Domain 3 (Benefits Realization, 26%) covers IT performance and oversight and management of IT-enabled investments, including business-case development and evaluation, IT investment management and reporting, performance metrics, and benefit evaluation methods (Domain 3B1–3B4). Domain 4 (Risk Optimization, 19%) covers risk frameworks, enterprise risk management, risk appetite and tolerance, and risk assessment methods (Domains 4A–4B). Supporting tasks include ensuring that a business-case development and benefits-realization process for IT-enabled investments has been established (Task 5); that policies and standards inform decision-making on IT-enabled business investments (Task 14); that IT initiatives are prioritized (Task 21); that IT-enabled investments are managed through their economic lifecycle (Task 28); that ownership and accountability for those investments are assigned (Task 29); that benefits realization of IT-enabled investments, processes, and services is evaluated (Task 31); that a performance-management program is established (Task 32); and that improvement initiatives are based on results derived from performance measures (Task 33) (ISACA Certification Exam Candidate Guide, Appendix D). CGEIT does not name enterprise AI assistants, license expansion, hiring freezes, or AI-attributed layoffs. It does require, as professional practice, that IT-enabled investment decisions be tied to business cases, measured performance, benefits evaluation, resource optimization, and assigned accountability.

COBIT 2019 [11] places those duties on the governing body in the Evaluate, Direct and Monitor (EDM) domain. EDM02 (Ensured Benefits Delivery) exists to “secure optimal value from I&T-enabled initiatives, services and assets” and “a reliable and accurate picture of costs and likely benefits,” with an example alignment metric of the percent of I&T-enabled investments for which claimed benefits in the business case are met or exceeded (ISACA, [11] Framework: Governance

and Management Objectives, EDM02 Purpose and AG03). Practices include establishing the target investment mix using cost, strategy alignment, benefit type, risk, and expected ROI over the full economic life cycle (EDM02.01); evaluating and directing value optimization (EDM02.02–EDM02.03); and monitoring value optimization (EDM02.04). EDM03 (Ensured Risk Optimization) assigns the governing body evaluation, direction, and monitoring of I&T-related risk. EDM04 (Ensured Resource Optimization) requires that adequate business and I&T resources—people, process, and technology—be available “at optimal cost,” and EDM04.01 calls for continual examination of current and future financial and human resource needs and allocation principles (EDM04 Purpose and EDM04.01). EDM01 sets and maintains the governance system; EDM05 covers stakeholder engagement. COBIT is a voluntary framework. It supports benefit-realization review, risk assessment, resource optimization, monitoring, and documentation before and after I&T investments. It does not prescribe a fixed waiting period, and it does not classify AI-attributed layoffs.

NIST AI Risk Management Framework 1.0 is voluntary, non-sector-specific guidance [29], January 2023, doi:10.6028/NIST.AI.100-1. The Core functions are GOVERN, MAP, MEASURE, and MANAGE. GOVERN is cross-cutting. MAP establishes context and intended use. MEASURE employs quantitative, qualitative, or mixed-method tools “to analyze, assess, benchmark, and monitor AI risk and related impacts.” NIST states: “AI systems should be tested before their deployment and regularly while in operation” [29], §5.3. MEASURE 2.3 requires that performance or assurance criteria be measured and documented for conditions similar to deployment; MEASURE 2.4 requires monitoring of functionality and behavior in production; MEASURE 2.5 requires demonstration that the system to be deployed is valid and reliable, with documented limits on generalizability [29], MEASURE 2. The NIST AI RMF Playbook [67] operationalizes human oversight, audit logs, measurement histories, downstream actions, overrides, errors, complaints, escalation, and documented “go/no-go” decisions that remain independent of stakeholders’ vested financial or reputational interests (NIST, [67], GOVERN/MAP/MEASURE/MANAGE suggested actions, including MANAGE go/no-go language and MEASURE instrumentation of audit logs and override statistics; https://airc.nist.gov/docs/AI_RM_F_Playbook.pdf). NIST does not require a 3-, 6-, or 12-month post-deployment wait before

budget, scale-up, or workforce decisions.

GAO's Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities [36], 30 June 2021, <https://www.gao.gov/products/gao-21-519sp> organizes key practices under Governance, Data, Performance, and Monitoring. Governance practices include defining clear goals and objectives that translate into system requirements and performance metrics, assigning roles, engaging stakeholders, and documenting specifications [36], Framework Principle 1; organizational-level practices. Performance practices include defining precise, consistent, and reproducible metrics; assessing components and systems against those metrics; documenting methods; and defining procedures for human supervision (Framework Principle 3, practices 3.2, 3.6, 3.9). Monitoring practices address continuous monitoring of performance and assessing sustainment and expanded use (Framework Principle 4). GAO later applied related accountability questions in agency work, including generative-AI use-case management and workforce-capability findings [37]; [59]. GAO's accountability framework [36] is not a statute. It does not specify a fixed waiting period before license expansion or layoffs, and it does not create an official taxonomy of AI-caused versus AI-related workforce reductions.

OMB memoranda bind U.S. federal executive agencies, not private firms. OMB memorandum M-24-10 [68] (28 March 2024, Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence) required designation of a Chief AI Officer, AI use-case inventory reporting, an enterprise AI strategy for CFO Act agencies, and minimum practices for safety-impacting and rights-impacting AI, including an AI impact assessment, testing in a real-world context before relying on the system, ongoing monitoring, periodic human review at least annually and after significant modifications, additional human oversight for consequential actions, documentation, and—where benefits do not meaningfully outweigh risks—a decision not to use the AI [68], §§1, 5(c). In enumerating factors for identifying covered AI, M-24-10 [68] explicitly lists worker well-being among impacts to consider [68], §5(b). Those provisions are federal-agency requirements under the memorandum's then-operative text; they are not legal requirements for private U.S. organizations. OMB [38] (3 April 2025) rescinds and replaces [68] and redirects agencies toward accelerated adoption with minimum risk-management practices for "high-impact AI," discontinuation when such AI is not performing at an appropriate level, Chief AI Officer designation, public AI strategies (CFO Act agencies, within 180 days), continuous monitoring in AI-enabling infrastructure, and workforce capacity planning [38], Overview and Appendix §§2–4; [https://www.whitehouse.gov/wp-content/uploads/2025/02/\[38\]-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf](https://www.whitehouse.gov/wp-content/uploads/2025/02/[38]-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf). Neither memorandum specifies a 3-, 6-, or 12-month waiting period before budget expansion, enterprise licensing, or workforce restructuring. M-25-21's 180-day deadline [38] clock is a strategy-publication deadline for CFO Act agencies, not a post-deployment decision gate.

4.15.2 Fixed waiting periods: finding: No authoritative source examined for this subsection—CGEIT exam content and supporting tasks, COBIT EDM01–EDM05 [11], NIST AI RMF 1.0 [29] and the AI RMF Playbook [67], GAO reports [36], [37], [59], OMB M-24-10 [68], or OMB M-25-21 [38]—specifies a fixed waiting period of 3 months, 6 months, or 12 months before an organization may make major post-deployment decisions on budget, license expansion, production scale-up, hiring freezes, role elimination, or layoffs. Where time is mentioned, it is context-specific (for example, M-24-10's at-least-annual human review for covered federal AI [68]; M-25-21's 180-day federal strategy deadline [38]). This review therefore does not attribute a 3-/6-/12-month rule to CGEIT, COBIT, NIST, GAO, or OMB.

4.15.3 Workforce attribution and budget displacement: U.S. official workforce measurement remains at the firm-share level: Bonney et al. [14] report AI-related employment decreases in 2% of firms and augmentation-only use among 66% of AI-using firms. That 2% is not an employee-share layoff rate and is not a causal identification of Copilot-class tools. GAO [59] documents the IRS RAAS unit's loss of 63 employees who had supported AI amid 2025 staffing reductions and does not attribute those losses to AI substitution; GAO treats the skills loss as a risk to future AI success (causality code 4 in this review). BLS 2024–34 projections incorporate assumed AI productivity effects; they are forecasts, not measured enterprise-assistant impacts [69]. The Worker Adjustment and Retraining Notification Act requires covered employers to provide 60 days' notice of qualifying plant closings or mass layoffs [70]. WARN is general employment-law notice, not an AI-evidence gate and not a requirement to prove AI causation.

No official CGEIT, COBIT, NIST, GAO, or OMB category distinguishes "AI-enabled," "AI-related," and "directly AI-caused" workforce reductions as formal labels that organizations must use. This review therefore retains its existing causality codes (0–5) and, as an analytical proposal of this paper—not an official taxonomy—distinguishes five workforce/budget mechanisms: (1) direct AI substitution, in which AI replaces documented tasks or roles; (2) AI-contributing restructuring, in which AI is one factor among several; (3) AI-investment-induced budget displacement, in which resources are shifted toward AI and staffing or other programs are reduced as a consequence of that reallocation rather than because AI automated the work; (4) general cost-cutting or restructuring for broader financial reasons; and (5) temporal association only, in which layoffs follow AI adoption but causation is not established. Category (3) is not an existing CGEIT, GAO, or NIST class. It is proposed here because CGEIT Domain 2 and COBIT EDM04 require resource optimization across people and technology, CGEIT Task 31 and COBIT EDM02 require evaluation of claimed versus realized benefits, and the included evidence contains no independently verified ledger in which AI invoices are paired with subsequent layoffs (Section 4.13; $n = 0$ complete pairs). Gartner's finding that workforce-reduction rates were nearly equal across ROI bands [24] is consistent with the possibility of budget or strategy-driven cuts that are not benefit-justified; it does not prove mechanism (3) in any named firm.

4.15.4 Decision-stage model and minimum evidence package (proposal of this review): The following decision-stage model is a proposed governance gate derived from the evidence gap in this review and from the framework-supported principles above. It is not an existing formal requirement of CGEIT, COBIT, NIST, GAO, or OMB. The stages are: initial investment → pilot → production or scale-up → renewal or budget expansion → workforce restructuring or layoffs. Irreversible or high-impact decisions—especially layoffs, major budget changes, or enterprise-wide scaling—should, under this proposal, require sufficient post-deployment evidence on actual benefits, costs, risks, reliability, adoption, and workforce effects before authorization. Claimed productivity, modeled ROI, or vendor automation equivalents are not, by themselves, that evidence (Sections 4.6).

A governance body (board risk or technology committee, public-sector equivalent, or an accountable executive with documented delegation) should, under this proposal, review a minimum evidence package that is staged rather than uniform. Before initial investment: business need; pre-AI baseline on the intended operational metric; alternatives analysis (including non-AI process change); expected total cost of ownership (licenses, qualifying suite, consumption, integration, training, specialist staff, and security review); expected benefits with a measurement plan; predefined KPIs; preliminary cybersecurity/privacy and human-oversight design; an accountable decision owner; and predefined scale/modify/stop criteria. Before production or scale-up, and before enterprise-wide license expansion or renewal: the initial package plus realized (not only expected) benefits against the pre-AI baseline; adoption and utilization among intended users; reliability and quality (including nulls and error or rework); cybersecurity/privacy incidents or residual risk; documented human-oversight operation (overrides, escalation, fallback); a monitoring plan; and a go/no-go record independent of vendor commercial interest. Before irreversible workforce decisions (hiring freeze, role elimination, or layoff attributed to AI): the scale-up package plus a workforce-impact assessment that states the proposed mechanism (direct substitution, contributing factor, investment-induced budget displacement, general cost-cutting, or temporal association only); the evidence for that attribution; whether saved time was observed to become output, quality, slack, or headcount reduction; and human review by an official not solely responsible for the AI business case. If attribution cannot be documented, the decision should not be described as AI-caused.

4.15.5 Governance gap this paper addresses: Authoritative sources support predefined metrics, benefit-realization review, risk assessment, resource optimization, ongoing monitoring, post-deployment evaluation, human oversight, documentation, and accountability for IT- or AI-enabled systems. They do not supply a ready-made, AI-assistant-specific gate that tells a private or public organization when claimed minutes saved or a commissioned ROI model may authorize enterprise-wide licensing or layoffs, and they do not impose a fixed 3-/6-/12-month clock. The included empirical set shows the operational consequence of that gap: independently measured financial ROI for named enterprise assistants was not identified (Section 4.6); paired spend-and-layoff ledgers are absent

(Section 4.13); and organization-share reduction statistics are easily misread as employee-share displacement (Section 4.4). Tables XVIII and XIX separate official requirements and framework-supported principles from this paper's proposals. The longitudinal chain in Section 4.13.1 is why the proposed gate refuses the label "AI-caused" when only earlier links (announcement, availability, or same-year timing) are filled.

Section 4.28 tests, against the full included and complementary record, whether a CGEIT/COBIT-style decision maker actually receives the inputs those authorities ask for. That test is an evidence-adequacy finding. It is not a finding that CGEIT fails.

4.16 Vendor-Reported Evidence — Excluded from Independent Synthesis

The following records remain in the corpus for non-UX questions (cost, telemetry design, or labeled claims) and do not influence independent satisfaction, usability, complexity, training, or cross-platform UX conclusions. Exclusion reason: Vendor-affiliated evidence — excluded from independent user-experience synthesis. Table XX lists these records.

4.16.1 Enterprise agents as action-taking systems: The central organizational issue with agents is not whether they generate useful text. It is whether they can reliably and safely execute consequential business actions across enterprise systems—record updates, workflow execution, ticket changes, procurement steps, IT operations, or compliance writes. Assistant tasks (drafting, summarization, search, recommendation, code suggestion) leave the human as the actor of record. Agentic tasks (multi-step planning, tool calling, orchestration, case handling, ITSM actions) can change systems of record. Demo, pilot, production deployment, and independently evaluated production are different stages. A vendor keynote is not the last of those. Section 4.22 compares the two categories without pooling their outcome numbers.

Independently evaluated production in this set is still mostly assistant-class: suggested replies (+15% issues per hour; S002), coding completions (S003–S004), Copilot drafting and search (S001; UK evaluations), and ambient notes into Epic (S009). Custom tool-using agents reach production in a success-biased practitioner sample, almost always with a human gate at or before ten steps (S053). Named commercial agent platforms—Copilot Studio, Agentforce, ServiceNow AI Agents, Gemini Enterprise, ChatGPT/Claude tool-calling—have vendor availability and, in some cases, vendor-named customers. This search did not verify an independent organizational-outcome RCT that they safely take high-impact writes. Copilot Studio's own documentation specifies a confused-deputy pattern if maker connector credentials are shared. CISA's 2026 advisory treats least privilege, dedicated agent identities, and human approval of high-impact actions as first-order controls, not maturity extras. Table XXI records task, autonomy, systems touched, human approval, stage, benefit, reliability, and risk.

4.17 RQ22 — Where agent building becomes hard

Creating a simple, constrained agent is not the stage with the strongest independent evidence of difficulty. The

TABLE XVIII

AUTHORITATIVE GOVERNANCE SOURCES VERSUS THIS PAPER’S PROPOSED EVIDENCE GATE (PART 1 OF 2): PRINCIPLE, DECISION-MAKER, CONTEXT, AND REQUIRED EVIDENCE. NO ROW OTHER THAN THE OMB ROWS IS A LEGAL DUTY FOR PRIVATE U.S. ORGANIZATIONS. NO SOURCE SPECIFIES A FIXED 3-/6-/12-MONTH WAITING PERIOD. CITATIONS AND REMAINING FIELDS ARE IN TABLE XIX.

Source	Exact principle or requirement	Decision-maker	Decision context	Evidence required
ISACA CGEIT outline (2020)	Business-case and benefits-realization process for IT-enabled investments; performance management; improvement based on measured performance	Enterprise IT governance roles (board/executive as specified)	IT-enabled investment, prioritization, economic lifecycle	Business case; performance metrics; benefit evaluation
ISACA [11] EDM02	Ensured benefits delivery: reliable picture of costs and likely benefits; claimed benefits met or exceeded	Governing body evaluates, directs, monitors	I&T investment mix, value optimization, portfolio stage gates	Cost, alignment, benefit type, risk, expected ROI over economic life cycle; monitoring of value
ISACA [11] EDM03–EDM04	Ensured risk optimization; ensured resource optimization (people, process, technology at optimal cost)	Governing body	Risk acceptance; financial and human resource allocation	Risk evaluation/direction/monitoring; current and future resource-need assessment
[29]	Test AI systems before deployment and regularly while operating; measure and document performance, validity, reliability, security, privacy, and impacts	Organizational AI risk roles under GOVERN	Design, development, deployment, and operation of AI systems	TEVV documentation; production monitoring; documented limitations
NIST [67]	Documented go/no-go decisions independent of vested financial/reputational interests; audit logs; override and downstream-action measurement	AI actors with documented roles	Deployment, iteration, decommissioning	Risk tolerances; measurement histories; human-oversight stats
[36]	Governance, Data, Performance, Monitoring: goals → metrics; human supervision; continuous monitoring; assess expanded use	Oversight body and management; auditors/assessors	Considering, selecting, implementing, and expanding AI	Documented goals, metrics, assessments, monitoring evidence
OMB [68] (rescinded)	CAIO; impact assessment; real-world testing; ongoing monitoring; at-least-annual human review; do not use AI if benefits do not meaningfully outweigh risks; consider worker well-being	Agency head and Chief AI Officer (federal)	Federal safety- and rights-impacting AI	Impact assessment; testing; monitoring; public inventory documentation
OMB [38] (current federal memo)	Rescinds [68]; CAIO; public AI strategy (CFO Act, 180 days); minimum practices for high-impact AI; discontinue if not performing; continuous monitoring; AI workforce capacity	Agency head and Chief AI Officer (federal)	Federal agency AI, especially high-impact AI	Strategy, inventory/reporting, risk practices proportionate to use
This review (proposed gate)	Evidence-gated authorization of irreversible or high-impact decisions	Board/risk committee or accountable executive	Scale-up, renewal/budget expansion, restructuring/layoffs	Staged minimum package: baseline, TCO, realized benefits, KPIs, adoption, reliability, cyber/privacy, oversight, workforce attribution, monitor, owner, stop criteria

hardness that matters organizationally is the jump from assistant text to reliable actions (Section 2.2). Measuring Agents in Production (S053) surveyed practitioners and interviewed 20 teams already running production or pilot agents. Among those successful cases, 70% used off-the-shelf models without weight tuning, 79% relied on manual prompt construction (sometimes exceeding 10,000 tokens), 68% executed at most ten steps before a human intervention, and 85% of detailed interview cases built custom applications rather than adopting a third-party agent framework. The top remaining challenge was reliability—especially evaluating correctness—not first-day scaffolding. That sample is biased toward systems that already reached users; it cannot be read as a success rate for all starts.

Difficulty concentrates later. Testing studies of real agent repositories (S054, S055) show an inversion of effort: deterministic tools and workflows absorb most tests, while the model plan body and prompts are barely tested. Three-quarters of MAP deployed-agent teams had no formal benchmark; 74% depended primarily on human evaluation. Enterprise knowledge and data are a second hard cluster: UK GDS found in-product agents weak at identifying exact source documents; CIO/KeyBanc coverage of Agentforce describes months of data-cleanup and Data Cloud pre-work before outputs stabilize; EchoLeak shows that RAG retrieval can pull a malicious email into a production Copilot context. Identity and connectors are a third cluster: Copilot Studio’s official documentation specifies that maker-provided connector credentials let every end user exercise the maker’s privileges, and that banning those credentials breaks autonomous/background agents. That

is product behavior, not a satisfaction study. CISA’s 2026 Five Eyes advisory treats least privilege, dedicated agent identities, short-lived credentials, and human approval of high-impact actions as first-order operational requirements. Fig. 17 maps all 23 stages. Multi-agent orchestration, IBM watsonx, SAP Joule, Oracle agents, ServiceNow AI Agents, Vertex Agent Builder, and Bedrock Agents still lack independent named-org builder studies in this set: Insufficient comparable independent evidence.

4.18 RQ23 — Cybersecurity complexity (classes A–E)

Class A (theoretical) and class B (laboratory) evidence is strong: NIST AI 100-2 and AI 600-1 define prompt injection and poisoning; DeBenedetti et al. [41] report benign-task utility below 66% even without attack; NIST/CAISI [32] reports attack-success rates rising from an 11% strongest baseline to 81% on a Workspace hold-out. These are not organizational breaches.

Class C (vulnerability in a real enterprise product): EchoLeak [34] demonstrated zero-click indirect prompt injection against Microsoft 365 Copilot—malicious email, XPIA bypass, Graph-context leak, markdown/Teams-proxy exfiltration. Microsoft issued a critical advisory and a May 2025 server-side fix and stated there was no evidence of customer exploitation. We do not describe EchoLeak as a confirmed organizational incident (class E).

Those classes matter because agentic tasks, unlike summarization, can change tickets, files, or entitlements (Section 2.2). Class D (real deployment problems): maker-credential

TABLE XIX

AUTHORITATIVE GOVERNANCE SOURCES VERSUS THIS PAPER’S PROPOSED EVIDENCE GATE (PART 2 OF 2): TIMING, OVERSIGHT, DECISION RELEVANCE, STATUS, AND EXACT CITATION. COMPANION TO TABLE XVIII.

Source	Pre- or post-deployment	Timing requirement	Human oversight	Budget / scaling / workforce relevance	Status and citation
ISACA CGEIT outline (2020)	Both (lifecycle)	None specified (no 3/6/12-month wait)	Not AI-specific; accountability assigned (Tasks 10, 29)	Supports evidence before treating claimed benefits as realized; no AI-layoff rule	Professional practice / certification content (not law). ISACA CGEIT Examination Content Outline, Domains 2–4; Candidate Guide Appendix D, Tasks 5, 14, 21, 28–33. https://www.isaca.org/credentialing/cgeit/cgeit-exam-content-outline
ISACA [11] EDM02	Both	None specified	Governing-body monitoring, not an AI HITL rule	Directly relevant to budget and scale-up; no layoff taxonomy	Voluntary framework principle. ISACA, [11] Framework: Governance and Management Objectives, EDM02 Purpose, AG03, EDM02.01–EDM02.04
ISACA [11] EDM03–EDM04	Both	None specified	Not specified as model-level HITL	Supports reviewing whether AI spend displaces staff or other resources; does not name that category	Voluntary framework principle. [11], EDM03; EDM04 Purpose and EDM04.01
[29]	Before deployment and during operation	“Regularly” — no 3/6/12-month decision wait	GOVERN/MAP human-oversight documentation	Supports evidence before scale-up; no budget or layoff mandate	Voluntary guidance. [29] (2023), §5.3 and MEASURE 2.3–2.5. https://doi.org/10.6028/NIST.AI.100-1
NIST [67]	Throughout lifecycle	None specified	Explicit human oversight, appeal/override, escalation	Supports refusing scale-up when measurement is absent; not a layoff rule	Voluntary suggested actions. NIST [67], MANAGE go/no-go; MEASURE audit-log and override actions. https://airc.nist.gov/docs/AI_RMF_Playbook
[36]	Before implementation and for sustainment/expansion	None specified	Practice 3.9: procedures for human supervision	Supports evidence before expanded use; no AI-layoff attribution rule	Accountability framework (not statute). [36] (30 June 2021), Principles 1, 3 (incl. 3.2, 3.6, 3.9), 4. https://www.gao.gov/products/gao-21-519sp
OMB [68] (rescinded)	Pre-use minimum practices and ongoing	Human review at least annually (not a 3/6/12-month layoff wait)	Required additional human oversight for consequential actions	Federal only; worker well-being is a risk factor, not a private-sector layoff rule	Then-operative federal requirement; not applicable to private firms. OMB [68] (28 Mar. 2024), §§1, 5(b)–5(c)
OMB [38] (current)	Both	180-day strategy deadline — not a post-deployment decision wait	Policies for consequential decision-making; cease use if mitigation impossible	Federal only; maximize value of existing investments; not a private layoff rule	Current federal memorandum; not a private-sector legal duty. OMB [38] (3 Apr. 2025), Overview and Appendix
This review (proposed gate)	Stricter after deployment for scale-up and workforce actions	No fixed clock proposed as a framework rule; sufficiency of evidence, not elapsed months	Required human review of attribution and go/no-go	Directly addresses budget, enterprise licensing, and layoff claims	Paper proposal. This manuscript, §4.15.1; grounded in CGEIT Tasks 31–33, COBIT EDM02/EDM04, NIST MEASURE, GAO Principles 3–4

TABLE XX

VENDOR-AFFILIATED OR UNCLEAR RECORDS RETAINED FOR NON-UX QUESTIONS. THESE ROWS DO NOT ENTER INDEPENDENT SATISFACTION, USABILITY, TRAINING, OR BUILDER-EXPERIENCE CONCLUSIONS.

ID	Product	Why excluded from primary UX	What it claimed (not synthesized)
S001 / S045	M365 Copilot	Microsoft authors / MSR / WTI	Email-time RCT; vendor survey items
S003 / S004	GitHub Copilot	Microsoft-affiliated authors / vendor study	Developer speed/tasks
S028	M365 Copilot	Microsoft partner (Cloudforce) ran the pilot	82% daily use; 38 min/day (n=85)
S029	ChatGPT Edu	UNCLEAR independence	NPS 55 / 64% promoters
S031 / S046	ChatGPT Enterprise	OpenAI authors / vendor report	Usage growth; 40–60 min/day claims
S041	M365 Copilot	Forrester TEI commissioned by Microsoft [28]	116% ROI; modeled training cost
S044	M365 Copilot	Vendor price list	List \$30/user/month
S047	Klarna assistant	Company PR	CSAT on par; \$40M claim
S052	M365 Copilot	UNCLEAR (Microsoft-led Prompt-A-Thons)	Time-saving narrative; oversharing risk

TABLE XXI
ASSISTANT VERSUS AGENTIC TASKS. FULL ROWS ARE IN THE LONGITUDINAL WORKBOOK. VENDOR DEMONSTRATIONS ARE NOT PRODUCTION PERFORMANCE.

Task	Platform	Autonomy / HITL	Stage	Measured benefit	Main risk
Suggested CS replies	GPT assistant	Low / human sends	Indep. eval. production	+15% IPH (S002)	Not a CRM write-agent
Code suggestion	GitHub Copilot	Low / human commits	Indep. experiments	+26% tasks (S003)	IP; not autonomous merge
Draft/search	M365 Copilot	User-initiated Graph	Licenses + UK evals + RCT	Email -12% ITT; mixed output	Permission inheritance; EchoLeak C
Ambient note	Abridge/Epic	Draft; clinician signs	U.S. RCT production	Well-being (S009)	PHI if review skipped
Multi-step tools	Mostly custom (MAP)	≤10 steps then human	Prod/pilot survey	Self-report time; no \$ ROI	Evaluation debt
Case handling	Agentforce	Vendor-claimed	Vendor-claimed production	Not independently measured	CRM writes; data foundations
ITSM / ops	ServiceNow AI Agents	Vendor-claimed	Vendor availability	Not independently measured	Ticket/CMDB writes
Low-code agents	Copilot Studio	Can be autonomous	Spec + paid SKU	Insufficient independent	Maker credentials = deputy
Workspace agents	Gemini Enterprise	Vendor-claimed	Launch/list Oct 2025	Not independently measured	Workspace access
High-impact write	Any LEVEL 3–4	High	Lab + advisory only	Insufficient evidence	Injection; unapproved action

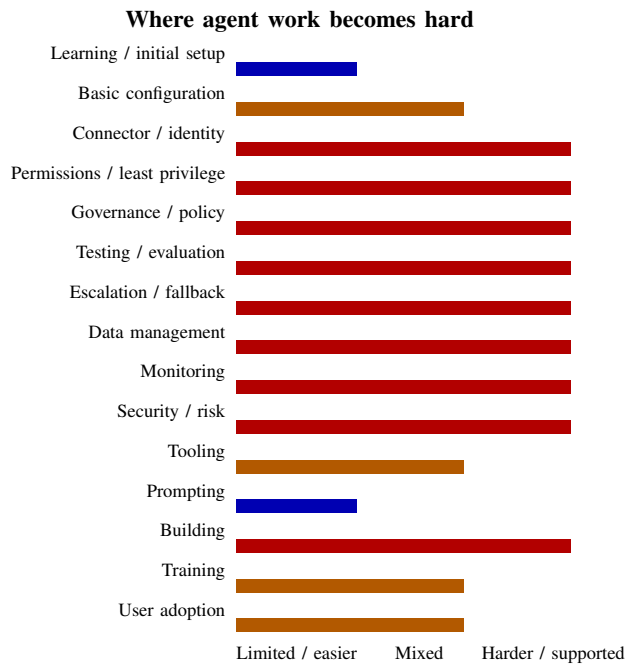


Fig. 17. Where independently documented difficulty concentrates across agent-building stages. Codes combine independent empirical studies, official advisories, and specified product behavior. The figure is not a scored ranking of vendors.

oversharing and autonomous-auth conflict (Copilot Studio specification); permission inheritance and classification limits (UK GDS/HMRC); federal policy/privacy drag (GAO); CISA’s finding that most IAM platforms were not designed for agent-class principals. Class E: this search did not verify a confirmed production breach caused by an enterprise AI agent. Contemporaneous reporting on the January 2026 Step Finance treasury loss attributes compromise to executive devices and keys, not to confirmed agent execution; later commentaries that relabel that event as an agent incident were not used.

4.19 RQ24 — Custom-coded versus platform-built agents

Independent empirical evidence on builders is thicker for custom-coded agents than for named low-code suites. MAP: most successful interviewed teams wrote custom applications and then deliberately limited autonomy. OSS testing papers document evaluation debt in those stacks. Low-code does not

turn an action-taking agent into a harmless assistant (Section 2.2). Low-code platforms are not shown, independently, to delete authentication, authorization, RAG grounding, evaluation, or monitoring. Copilot Studio documentation shows those problems moved into connector identity, environment DLP, and admin policy—the confused-deputy pattern is a configuration failure mode, not a missing Python file. Agentforce coverage locates delay in data foundations rather than in the first agent sketch. We do not conclude that custom is harder overall or that low-code is safer. We conclude that low-code does not, on present independent evidence, remove enterprise complexities; it relocates several of them into configuration and governance. Custom development adds infrastructure, state, observability, and test-harness work (MAP; testing studies). Both paths still require humans for evaluation and high-impact actions.

4.20 RQ25 — Prototype-to-production gap

A working demo of an agentic write is not independently evaluated production (Section 2.2). Greatest independently documented friction is not “write the first prompt.” It is the interval after a demo works: connecting live inconsistent data and APIs; proving evaluation; passing identity/security review; logging enough to assign accountability; and keeping quality after scale. NANDA’s preliminary custom-tool stall, Deloitte’s scale-up failure share, Gartner’s prediction that over 40% of agentic projects will be canceled by end-2027 (cost, unclear value, inadequate risk controls—a prediction, not an observed rate), Agentforce data-prep delays, and Klarna’s quality-driven human return are the available independent or labeled-claim traces. MAP’s successful teams already chose short workflows and human gates—consistent with surviving that interval by reducing autonomy rather than by proving full autonomy. Unverified 88–89% “fail before production” blog statistics were excluded because primary methods could not be checked.

4.21 RQ26 and hidden human labor / true cost

Does adoption eliminate organizational labor or redistribute it? The evidence does not support a computed net FTE. Where agents reach production, humans remain: 74% human evaluation (S053); 45% of merged Claude Code PRs revised by people (S060); continuous oversight in UK Copilot; 2025

TABLE XXII

SYNTHESIS OF COMPLEXITY, CYBERSECURITY-CLASS, LABOR-SHIFT, AND NET-VALUE QUESTIONS. “INSUFFICIENT EVIDENCE” IS AN EXTRACTION RESULT, NOT A CLAIM THAT THE TRUE QUANTITY IS ZERO.

Question	Independent answer	Gap
Where is the hard part?	Integration, identity/permissions, evaluation/testing, governance, production/maintenance, cybersecurity—not first-prompt authoring among successful teams	Named low-code builder surveys still thin
Custom vs low-code	Custom adds infra/test debt; low-code moves auth/data/governance into config	No paired trial of both paths
Class E agent breach?	None verified	Incident under-reporting possible
Net labor / net value?	insufficient evidence	No comparable FTE or TCO panels
Eliminate or shift labor?	Consistent with SHIFT plus residual automation of tasks; not a computed identity	Need matched before/after staffing

Hidden human labor — evidence map

Lost human labor may reduce future AI capacity • Loss of analytics and AI-support staff • Loss of specialist knowledge and expertise • Reduced internal capability to support AI systems • Net FTE effect remains unknown	New human work is still required • Human evaluation and review • Continuous oversight and revision • Identity and access administration • Security monitoring and red teaming • Maintenance and escalation
Net Human Labor Effect = UNKNOWN Available evidence does not support a defensible net FTE calculation. Agent deployment may reduce some tasks while simultaneously creating evaluation, oversight, security, identity, maintenance, and specialist-support work. This review therefore records <i>Net Human Labor Effect</i> , <i>Total Agent Cost</i> , <i>Realized Benefits</i> , and <i>Net Value</i> as insufficient evidence rather than estimating unsupported totals.	

Fig. 18. Residual and hidden human labor around production agents, including evaluation, revision, oversight, identity administration, and complementary specialist staff. No net FTE or net-value total is computed. Case-level coding is reported in Enterprise_Agent_Complexity_Cases.xlsx.

human CS reinvestment at Klarna after a claimed 700 FTE-equivalent. GAO’s IRS case shows the complementary labor most clearly: losing 63 AI-supporting staff is treated as a threat to future AI success, not as an automation dividend. CISA’s controls imply new roles—IAM for agents, approvers, monitors, red-teamers. Training hours, consultant invoices, and pre-production spend are almost never published with outcomes. Therefore Net Human Labor Effect, Total Agent Cost, Realized Benefits, and Net Value are recorded as insufficient evidence in every complexity-case row. Apparent workforce savings are not shown to remain economically meaningful after technical, security, governance, maintenance, and failed-implementation costs. That is an evidence gap, not a finding that savings are zero. Fig. 18 maps residual human roles without inventing a net FTE. Table XXII synthesizes the complexity, cybersecurity-class, labor-shift, and net-value questions.

How much cost is incurred before an organization knows whether an agent will reach production and create measurable value? Independent ledgers pairing proof-of-concept, security-review, rework, and maintenance dollars with a binary production outcome were not identified. Qualitative evidence says the unknown interval is long (months of data work; security review without an audit trail; custom tools that never leave pilot). A working prototype, completed vendor training, or a demo agent is not classified here as a successful enterprise

deployment or as positive ROI.

4.22 Assistive AI versus agentic AI — comparative synthesis

Existing results are preserved and not re-averaged. Each included deployment was re-read for what the system actually did. Suggested replies, Copilot mail/file drafting, GitHub completions, and ambient notes that a clinician signs are Category 1 even when the vendor says “AI.” MAP custom tool-using workflows, Copilot Studio autonomous/background agents, Agentforce service agents that can change CRM cases, and ServiceNow agents that can change ITSM records are Category 2 when those write-paths are in scope. Klarna’s chat handling is coded as a claimed high-volume assistant/agent hybrid and is not used as an independent Category 2 RCT.

On that coding, independent organizational evidence is currently more mature for assistive AI than for agentic AI, while autonomy, implementation complexity, security exposure, and governance demands increase as systems move from assistance toward action. That sentence is a synthesis of the included set, not a hypothesis imposed on it. Category 1 has multiple Tier A operational studies and independent public-sector user evaluations. Category 2 has practitioner production evidence that is success-biased (S053), laboratory security and reliability results, vendor general-availability claims, and official advisories—but this search did not identify an independent organizational-outcome RCT for a named commercial agent platform. Independently measured financial ROI was not identified in either category. Workforce class A (direct AI-caused layoff) was not assigned to either category. Table XXIII and Fig. 19 compare dimensions without merging incomparable numbers (for example, +15% issues per hour is not an Agentforce result).

4.23 Disclosure and comparability of public evidence

This subsection tests a reporting question, not a predetermined verdict. For each major platform class and each major survey, we asked whether the public source states deployment scale, active use, task type, autonomy, human intervention, organizational cost or TCO, implementation/training/security cost, measured productivity, quality, realized financial ROI, reliability or failure rates, security incidents, and workforce or hiring effects. Answers are Yes, Partial, or No against the included record. Employee self-report, executive expectation,

TABLE XXIII
ASSISTIVE VERSUS AGENTIC EVIDENCE. CLASSIFICATION IS BY BEHAVIOR AND AUTONOMY, NOT BRAND. FULL NOTES ARE IN
ENTERPRISE_AI_LONGITUDINAL_EVIDENCE.XLSX.

Dimension	Assistants / copilots	Agents / agentic systems	Evidence strength	Key limitation
Productivity	Local causal gains and documented nulls (S001–S003; DBT)	No named-platform org RCT; MAP self-report in success-biased sample	Stronger for assistants	Do not transfer % gains to agents
Satisfaction	High keep/satisfy in independent Copilot evals; mixed by task	Insufficient independent builder UX	Stronger for assistants	Vendor UX excluded from primary synthesis
Implementation	License/enabement plus classification and change management	Later-stage: data, eval, identity, logging, scale	Both documented; different objects	No paired build-path trial
Training	Hours rarely published	Human evaluation labor; TCO unpublished	Both insufficient as \$ lines	Do not invent hours
Reliability	Jagged by task; some quality reversals	Top MAP challenge; lab utility low under attack	Field vs survey/lab	Lab $\neq$ incident rate
Security	Inheritance; EchoLeak class C, not class E	Larger tool surface; maker-credential deputy; CISA controls	Product vuln vs spec/lab/advisory	No verified class E agent breach
Oversight	Human is actor of record	Successful teams cap steps and keep a human gate	Consistent	Vendor HITL not independently audited
Production maturity	Paid SKUs; RCT + independent user evals	Vendor GA; some custom production (MAP)	Assistants more mature as evidence	Census 18% $\neq$ agent share
ROI	Independent \$ ROI not identified	Independent \$ ROI not identified; cuts $\neq$ higher ROI (survey)	Both lack \$ ROI	TEI is modeled
Workforce	No Copilot-caused layoff in independent evals	Org-share survey cuts (class B); no class A named agent	Neither is class A	Org-share $\neq$ employee-share
Governance	Use policy, measurement, classification	Plus identity, write-scope, approval, rollback, monitoring, eval	Advisories speak most to agents	Scale-up rule is this review's proposal

Category 1 — Assists

Draft / search / suggest
Human is actor of record
Tier A org outcomes exist
Independent Copilot UX exists
Independent \$ ROI: none

Category 2 — Acts

Plan / tool-call / write
System of record can change
No named-platform org RCT
MAP: constrained production
Class E breach: none verified

Classify by behavior and autonomy, not brand.

Fig. 19. Evidence-maturity comparison. Category 1 assists a human; Category 2 can act inside enterprise systems. No causal arrow from assistant gains to agent ROI.

vendor claim, analyst survey, administrative or telemetry data, causal field experiment, and official statistics are not treated as interchangeable.

No included source reports the full product-level set together. The closest Category 1 packages still omit invoices, realized \$ ROI, and layoff outcomes: Dillon and colleagues measure time use for 7,137 workers in 66 firms but not TCO, quality scores, failure rates, or headcount; UK departmental Copilot evaluations report licenses, keep-tool items, and mixed task usability but not independently verified financial return; [18] measure coding-task throughput but not enterprise TCO. ChatGPT Enterprise telemetry (S031/S046) reports organizational use growth with unpublished seats and no independent causal productivity or ROI. In this search, Claude Enterprise and Gemini Enterprise contribute availability and list-price information, not independent organizational outcome studies.

Named Category 2 platforms are thinner on the same grid.

Copilot Studio, Agentforce, and ServiceNow AI Agents have vendor specifications or named-customer claims and, in Studio's case, a documented confused-deputy authentication behavior. They do not have independent public figures for seats in use, actual TCO, measured productivity, quality, realized ROI, failure rates, audited HITL, or workforce effects. MAP (S053) is the densest independent agent row and still omits dollars, unbiased start-to-production rates, and FTE. Counted as Yes cells on Fig. 20's eight fields, assistant experiment/evaluation rows contain multiple Yes/Partial marks on use, productivity, and HITL; named Agentforce / ServiceNow / Studio / Gemini / Claude Enterprise rows are No on scale census, TCO, productivity, ROI, failure rate, and workforce. That is a demonstrated field-by-field difference, not a slogan.

Surveys are not a substitute product ledger and are not comparable to one another without restating units. Census BTOS measures firm-share adoption and a 2% AI-related

TABLE XXIV

PUBLIC DISCLOSURE BY PLATFORM OR EVIDENCE CLASS. Y = REPORTED; P = PARTIAL; N = NOT REPORTED IN THE INCLUDED PUBLIC SOURCE. FULL MISSING-DATA NOTES ARE IN ENTERPRISE_AI_LONGITUDINAL_EVIDENCE.XLSX. THIS TABLE IS NOT A PRODUCT RANKING.

Source	Scale	Use	TCO	Prod.	ROI \$	HITL	Workforce	Independence
M365 Copilot RCT (S001)	P	P	P	Y	N	Y	N	Vendor-affiliated
M365 Copilot UK evals	P	P	P	P	N	Y	N	Independent (intl)
GitHub Copilot RCTs	P	Y	N	Y	N	Y	N	Vendor-affiliated
ChatGPT Enterprise telemetry	P	P	N	N	N	P	N	Vendor-affiliated
Claude / Gemini Enterprise	N	N	N	N	N	N	N	Vendor/press
Copilot Studio / Agentforce / ServiceNow	N	N	N	N	N	P*	N	Vendor (*configurable)
MAP custom agents (S053)	P	Y	N	P	N	Y	N	Independent
Census BTOS any AI	Y (firm share)	P	N	N	N	N	P (2% firms)	Official

TABLE XXV

INCLUDED SURVEYS ARE FRAGMENTED: DIFFERENT POPULATIONS, UNITS, AND CONSTRUCTS. NONE MEASURES THE FULL ORGANIZATIONAL OUTCOME SET ON ONE NAMED PRODUCT.

Survey	n / frame	Adoption	ROI / \$	Workforce	Autonomy	Behavior type
Census BTOS (S014)	U.S. firms	Y	N	Y (firm share)	N	Official self-report
Bick et al. (S011)	U.S. households	Y	N	N	N	Household self-report
[52] (S037)	1,993	Y (resp. share)	Self-report	Partial	Partial	Executive self-report
Deloitte Wave 4 (S038)	2,773 pilots	Y (frame)	Partial	Partial	Partial	Executive / adopter
PwC Agent US (S040)	308 execs	Plans	N	Expectation	Partial	Executive expectation
Gartner n=350 (S032)	Agent-using \$1B+	Frame	Self-report bands	Org-share cuts	Frame	Executive self-report
Gartner CS 321 / CHRO 110	Leaders	Implied	N	Y (org share)	Partial	Leader self-report
MAP (S053)	86 + 20	Production sample	N \$	N FTE	Y (step caps)	Actual practice (success-biased)
Forrester TEI (S041) [28]	Composite	Assumed	Modeled 116%	N verified	N	Modeled / commissioned

employment-decrease item, not Copilot TCO. McKinsey's 88% is a respondent-share of functional use. Gartner's ~80% is an organization-share of some reduction among large agent-using firms. PwC's agent survey reports executive expectations of headcount and role change. The Forrester TEI [28] is a commissioned model. MAP records constrained production practice among successful teams. None of these instruments jointly measures adoption, satisfaction, productivity, realized ROI, workforce counts, agent autonomy, and security incidents on one named product. Table XXV records that directly: no survey in the included set is a complete organizational outcome instrument.

The pattern supported by that grid—not assumed in advance—is: the evidence base is fragmented and incompletely transparent. Comparable product-level data on actual cost, realized benefits, reliability, human intervention, and workforce effects are rarely reported together. Specific missing fields are listed in Table XXIV; the statement is those omissions, not a general accusation of concealment. Agentic-system public evidence is materially thinner than assistant/copilot evidence on independent productivity, independent user evaluation, and any joint cost–benefit–workforce packet (Section 4.22; zero named-platform organizational-outcome RCTs versus several Category 1 field experiments).

M365 RCT	P	P	P	Y	N	N	Y	N
UK Copilot	P	P	P	P	N	P	Y	N
GitHub RCT	P	Y	N	Y	N	P	Y	N
ChatGPT Ent.	P	P	N	N	N	N	P	N
Claude Ent.	N	N	N	N	N	N	N	N
Gemini Ent.	N	N	N	N	N	N	N	N
Studio	N	N	N	N	N	N	P	N
Agentforce	N	N	N	N	N	N	P	N
ServiceNow	N	N	N	N	N	N	P	N
MAP custom	P	Y	N	P	N	P	Y	N
	Scale	Use	TCO	Prod.	ROI \$	Fail %	HITL	WF

Fig. 20. Field-by-field disclosure (Yes / Partial / No). Named commercial agent platforms are emptier than Copilot-class experiment and evaluation rows. No row is a complete ledger.

4.24 Firm-level AI investment and later hiring or headcount

The complementary question is: did firms that increased AI investment more aggressively subsequently reduce hiring or headcount more than comparable firms, after controlling for firm and macroeconomic conditions? A valid causal panel was not constructible from the public record assembled for this review. Complete pairs of actual buyer AI expenditure and subsequent official headcount or hiring remain $n = 0$

TABLE XXVI

DESCRIPTIVE FIRM-PERIOD PANEL. SPEND TYPES ARE NEVER COMBINED. THE DiD/EVENT-STUDY SPECIFICATION THAT WAS PRE-SPECIFIED AND NOT RUN IS IN SUPPLEMENTARY S18. NO COEFFICIENT IS REPORTED.

Organization	Period	Spend object (type)	Hiring / HC note	Why not DiD-ready
Microsoft	FY23–25	PPE +\$28.1B → \$64.6B (actual vendor capex)	Supplier, not buyer HC	Wrong economic object
IBM	2023+	AI \$ not public	Class 4 hiring pause (planned)	No spend series; no controls
Salesforce	2023–25	Vendor product \$ ≠ buyer TCO	2023–24 cuts class 6/7; 2025 eng. freeze class 4	No buyer ledger; vendor interest
Bank of America	2024–26	~\$4B tech (not AI-only)	Attrition mix (class 5)	AI \$ not separable
Duolingo	2023–24	TCO unpublished	Contractor non-renewal	No \$ or quarterly counts
OneGov agencies	2025	Promo committed price	HC not in PRs	No paired employment
Dillon RCT firms	2023–24	300-license incentive	HC not an outcome	Anonymized; no HC
Complete buyer \$ + HC pair	2022–26	None identified	—	n = 0

(Section 4.13). Therefore no difference-in-differences or event-study coefficient, standard error, confidence interval, or effect size is reported. Forcing a regression on Microsoft's PPE additions, the \$30 Copilot list price, GSA promotional prices, or the Forrester modeled composite [28] would treat the wrong spending object as treatment and would not satisfy parallel-trends or comparability requirements.

What a descriptive firm-period panel can show (Table XXVI). Microsoft's additions to PPE rose from \$28.107 billion (FY2023) to \$44.477 billion (FY2024) to \$64.551 billion (FY2025). That is actual SEC cash investing by a supplier of AI infrastructure, not buyer Copilot TCO and not a customer headcount series. IBM's 2023 hiring pause has a class 4 management mechanism and no public AI-spend time series. Salesforce's 2025 engineer-hiring statement has claimed productivity and no published buyer-AI ledger; its 2023–24 reductions remain class 6/7. Bank of America's ~\$4 billion technology-initiatives language is not an AI-only ledger. Duolingo's contractor non-renewal has no public AI TCO. Federal OneGov prices are promotional. The Dillon RCT measures time use after a 300-license incentive and does not report headcount. IRS RAAS lost staff who supported AI and has no invoice in the GAO extract.

Fields that prevent estimation: (i) actual buyer AI spend, as opposed to vendor capex, list price, promo price, or modeled TEI; (ii) a quarterly or annual headcount or hiring series for the same named firms; (iii) a comparable set of firms that did not raise AI spend; (iv) job-posting extracts linked to those firms; (v) independently measured automation or productivity for the same firm-periods, without which spend → automation → headcount cannot be claimed even if a reduced form existed. Closest published designs in the literature are not this estimand: Westby et al. [61] use ChatGPT timing and junior software vacancies (~14–15% relative decline), and Brynjolfsson, Chandar, and Chen [62] use occupation AI-exposure and young-worker employment (-16% relative). Both are AI-contributing exposure or event designs, not firm AI-dollar treatments, and they do not authorize a spend–headcount elasticity from this review.

4.24.1 Supplier panel versus buyer panel: Panel A is major U.S.-relevant AI/cloud/software suppliers with public 10-K series: Microsoft, Alphabet, Amazon, Meta, Salesforce, ServiceNow, and IBM (Oracle and other filers were screened; they are included in the workbook only where a reliable headcount-plus-investment pair was keyed). Panel B is organizations with

documented procurement or deployment of named enterprise assistants/agents. The two panels are not pooled. A supplier's data-center capex is not a customer's Copilot invoice.

Panel A, descriptive (Table XXVII). Microsoft additions to PPE rose from \$23.886 billion (FY2022) to \$28.107 billion (FY2023, +17.7%), \$44.477 billion (FY2024, +58.2%), and \$64.551 billion (FY2025, +45.1%). Headcount was 221,000 in FY2022–FY2023, then 228,000 in FY2024–FY2025 (+3.2% then 0%). Capex intensity rose; net headcount did not fall. Alphabet employees fell from 190,234 (2022) to 182,502 (2023, -4.1%) while capex was roughly flat (~\$31.5B to ~\$32.3B), then capex rose to \$52.5 billion in 2024 (+~63%) while headcount was 183,323 (+0.4%). Meta's 2023 'Year of Efficiency' cut headcount 22.1% (86,482 to 67,317) while capex fell 12.3% (\$32.04B to \$28.10B); 2024 capex rose to \$39.23 billion and headcount rose 10.0% to 74,067. Amazon purchases of PPE were \$63.645 billion (2022), \$52.729 billion (2023, -17.2%), and \$82.999 billion (2024, +57.4%); year-end employees were about 1,541,000 (2022) and 1,556,000 (2024). Salesforce cash capex declined (\$798M / \$736M / \$658M) while the firm ran 2023–24 reductions and a 2025 engineer-hiring freeze; that capex series is not hyperscale AI infrastructure. None of these filings provides an AI-only dollar that can be used as treatment. The descriptive pattern does not show that larger subsequent capex increases produced larger subsequent headcount declines: Meta's big cut came with lower capex; Microsoft's capex boom came with flat-to-up headcount; Alphabet's 2023 cut preceded the 2024 capex jump.

Panel B, descriptive (Table XXVIII). Auditable buyer-side dollars are almost entirely promotional or incentive objects: GSA OneGov (\$1 per agency-year ChatGPT; Copilot zero-price year for qualifying G5); Dillon's 300-license incentive (anonymized firms, no headcount outcome). License counts without invoices (UK 20,000; Virginia Tech 177; CSIRO 300) are not TCO. Vendor-named ChatGPT Enterprise, Agentforce, ServiceNow, Studio, and Gemini Enterprise customers are coded expenditure UNKNOWN. Bank of America's ~\$4 billion technology-initiatives language is not an AI ledger. Duolingo and IRS have workforce mechanisms and UNKNOWN AI spend. A buyer DiD of AI \$ on later hiring is still not estimable (n of complete buyer spend–HC pairs = 0).

TABLE XXVII
AI SUPPLIERS. TOTAL CAPEX IS NOT TREATED AS AI-ONLY. FULL ROWS: ENTERPRISE_AI_LONGITUDINAL_EVIDENCE.XLSX.

Company	Year	Investment object (not AI-only unless split)	YoY inv.	Headcount	YoY HC	Quality
Microsoft	FY22–FY25	PPE additions \$23.9B → \$64.6B	+17.7% / +58.2% / +45.1% YoY	221k → 228k → 228k	0% / +3.2% / 0%	SEC; not Copilot TCO
Alphabet	2022–24	Capex ~\$31.5B → \$32.3B → \$52.5B	+~2% / +~63%	190,234 → 182,502 → 183,323	-4.1% / +0.4%	10-K; not AI-only
Amazon	2022–24	PPE purchases \$63.6B → \$52.7B → \$83.0B	-17% / +57%	1,541k (2022); 1,556k (2024)	Ops+corp mix	10-K; fulfillment+DC
Meta	2022–24	Capex+FL princ. \$32.0B → \$28.1B → \$39.2B	-12% / +40%	86,482 → 67,317 → 74,067	-22% / +10%	Earnings/10-K
Salesforce	FY23–25	Cash capex \$798M → \$658M	Down	76,453 at 31 Jan 2025	Prior cuts; 2025 eng. freeze	Capex ≠ AI \$

TABLE XXVIII
ENTERPRISE BUYERS/USERS. EXPENDITURE IS UNKNOWN UNLESS AUDITABLE. DO NOT INFER SPEND FROM 'USES AI.'

Buyer	Year	Auditable spend?	Deployment	HC / hiring	Quality
U.S. OneGov agencies	2025	Promo \$1/agency; Copilot \$0/12 mo G5	ChatGPT Ent.; M365 Copilot	Not in PRs	Committed promo
Dillon RCT firms	2023–24	300-license incentive only	M365 Copilot	HC not measured	Not TCO
UK GDS / VT / CSIRO	2024–25	Seats; invoices UNKNOWN	M365 Copilot	No Copilot layoff finding	Seats ≠ \$
Named ChatGPT Ent. logos	2023+	UNKNOWN	ChatGPT Enterprise	UNKNOWN	Use ≠ spend
Duolingo / BAC / IRS	2023–26	UNKNOWN as AI-only \$	Content AI / Erica / IRS AI	See Section 4.4.2	Mechanism ≠ ledger
Named agent-platform customers	2024–26	UNKNOWN	Vendor-claimed	UNKNOWN	Logos ≠ invoices

4.25 Administrative labor-market evidence (company versus industry/geography)

The complementary question in this subsection is whether stronger AI adoption or AI-investment growth is followed by weaker hiring, higher separations, slower payroll growth, more non-backfill, or lower headcount growth than in comparable firms or industries. Company evidence (Level 1) and Census/BLS administrative evidence (Level 2) are kept in separate files and then read against each other. Neither IRS nor Social Security employer tax microdata are used. Public QWI is an industry–geography–firm-size file; it is not a named-employer extract. A valid industry difference-in-differences or event study of AI spend or AI use on QWI/JOLTS/QCEW outcomes was not constructible: those public series have no AI-investment treatment column, and Census Business Trends and Outlook Survey AI-use items are firm-share survey data that cannot be merged to public QWI without a confidential linkage this review does not have. No industry-level coefficient is reported. Forcing a ChatGPT-calendar dummy onto national JOLTS openings would attribute the entire post-2022 cooling—including manufacturing and government—to enterprise assistants and is rejected.

Level 2, JOLTS (Table XXIX; Fig. 21). Using BLS annual not-seasonally-adjusted tables from the 2026 M01 release, U.S. job openings (annual average of monthly levels) fell from 11,178,000 in 2022 to 9,268,000 (2023, -17.1%), 7,660,000 (2024, -17.4%), and 7,089,000 (2025, -7.5%). Preliminary June 2026 openings were 7,359,000. Annual hires fell from 76,832,000 (2022) to 70,343,000, 64,515,000, and 62,972,000. Annual quits fell from 50,503,000 to 44,205,000, 39,285,000, and 38,029,000. Annual layoffs and discharges rose from 17,644,000 (2022) to 19,847,000, 20,014,000, and 21,239,000 (2025). Private-sector paths match the totals. The 2022 starting point is the post-pandemic tightness peak, not a pre-AI steady

state. Controls that any causal reading would have to confront therefore include overhiring and vacancy hoarding in 2021–22, the 2022–23 interest-rate cycle, industry demand, and ordinary recession risk—not only the commercialization of ChatGPT Enterprise (28 August 2023) or Microsoft 365 Copilot (1 November 2023).

Industry and geography slices remain aggregates. Information-sector hires fell from 1,270,000 (2022) to 874,000 (2023, -31.2%) and then were essentially flat (904,000 in 2024; 903,000 in 2025). Professional and business services hires fell from 14,943,000 (2022) to 12,533,000 and 11,791,000, then rose to 12,274,000 in 2025. Manufacturing hires and openings declined steadily through 2025. Federal hires fell from 541,000 (2023) to 427,000 (2024) and 321,000 (2025); federal openings fell from 163,000 to 142,000 and 108,000. Among regions, 2023 job openings declined in every Census region and most sharply in the West (-22.5%). Layoffs and discharges rose in every region from 2022 to 2025. These cells describe labor-market cooling and a rise in involuntary separations. They do not identify which employers adopted Copilot, Agentforce, or ChatGPT Enterprise, and they are not a WARN list.

Level 2, QWI and QCEW. Public Quarterly Workforce Indicators release employment, new hires, separations, job creation and destruction, payroll, and earnings by industry, geography, and firm-size class. The R2026Q1 vintage extends national and state files to about 2025Q1. Those are the correct fields for an industry/geography stock-and-flow picture. They are the wrong fields for a named-firm layoff claim. Establishment-level or EIN-level QWI is confidential LEHD and is not in this package. QCEW similarly measures establishment employment and wages by NAICS and county without an AI-spend item. A full QCEW NAICS-quarter panel was not re-keyed solely to run an invalid treatment regression. Together with the Current

TABLE XXIX

BLS JOLTS ANNUAL NSA LEVELS, 2026 M01 RELEASE (TABLES 15, 17, 21, 23). INDUSTRY AND GOVERNMENT ROWS ARE AGGREGATES—NOT NAMED-FIRM LAYOFFS AND NOT AN AI-CAUSE SERIES. JUNE 2026 OPENINGS 7,359,000 (P). FULL ROWS: WORKBOOK SHEET ADMIN_JOLTS. COMPLEMENTARY TO PRISMA $n = 61$.

JOLTS series (NSA)	2022	2023	2024	2025	2022–25
Openings, annual average (000s)	11,178	9,268	7,660	7,089	-36.6%
Hires, annual total (000s)	76,832	70,343	64,515	62,972	-18.0%
Quits, annual total (000s)	50,503	44,205	39,285	38,029	-24.7%
Layoffs & discharges (000s)	17,644	19,847	20,014	21,239	+20.4%
Information hires (000s)	1,270	874	904	903	-28.9%
Prof. & business svcs hires (000s)	14,943	12,533	11,791	12,274	-17.9%
Federal hires (000s)	502	541	427	321	-36.1% vs 2022

Employment Statistics payroll survey, the administrative stock series show that U.S. payroll employment did not collapse in parallel with the 2022–25 openings decline. JOLTS cooling is a tightness and turnover story first. It is not, by itself, a national displacement census.

Level 1, WARN and SEC (Table XXX). WARN and 8-K/10-K reductions that can be named—Meta 2022–23, Alphabet 2023, Amazon corporate 2022–23, Salesforce 2023–24 plus the 2025 engineer hiring freeze, IBM’s 2023 planned back-office pause, Duolingo contractor non-renewal—remain the company file already coded in Sections 4.4.2 and 4.24.1. A complete 2022–2026 national WARN extract flagged for AI was not constructed; state portals are fragmented and DOL does not publish an AI-reason national file. The 21.239 million JOLTS layoffs and discharges in 2025 are all-cause and must not be rewritten as that missing WARN census. SEC extracts continue to show mixed headcount beside rising, falling, or unsplit capex (Table XXVII). Personnel expense was not extracted as a comparable single line across filers. Management language about AI replacing or slowing hiring exists for IBM and Salesforce engineers and as a mixed digital/AI attrition story at Bank of America; those statements are Level 1 speech acts, not QWI cells.

Triangulation (Table XXXI), without causal upgrade. (i) Named AI-supplier capex and named headcount do not move as a simple inverse (Section 4.24). (ii) National openings, hires, and quits did fall, and layoffs and discharges did rise, over the same years that enterprise assistants became purchasable. That is temporal association at industry scale after a historically tight 2022. (iii) Information hiring’s 2023 step-down matches the documented tech reset more closely than it matches Copilot general availability (November 2023). (iv) Federal hiring and openings weakened in 2024–25 alongside public-sector staffing actions (IRS RAAS is class 6, not assistant substitution). (v) Census BTOS still reports that only 2% of firms claimed an AI-related employment decrease and that 66% of using firms were augment-only—firm shares, not JOLTS flows, and not a contradiction of either series if units are kept straight. The review does not write that AI caused the JOLTS decline, and it does not write that AI caused any named layoff unless a mechanism is documented. Explicit class 1 completed FTE substitution with a named-agent invoice remains $n = 0$.

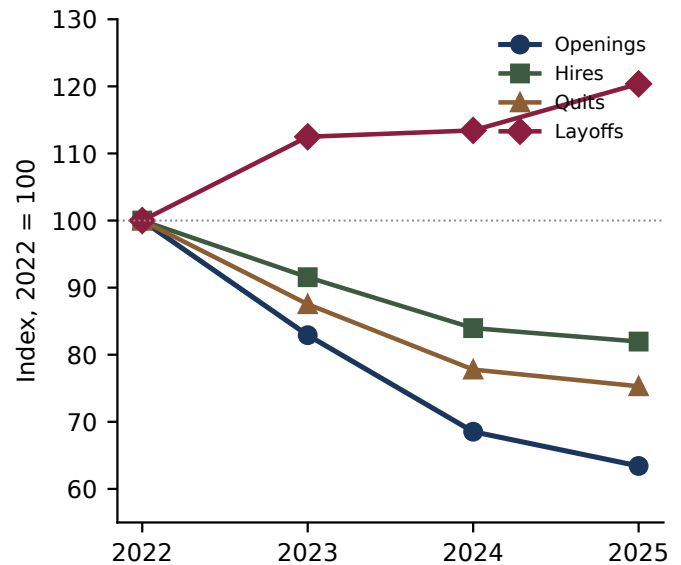


Fig. 21. JOLTS openings, hires, quits, and layoffs/discharges indexed to 2022 = 100. Administrative industry totals. Do not treat the post-2022 openings decline as AI-caused named-employer displacement.

4.26 Complete AI workforce-substitution cases — targeted search

The complementary question is whether any named organization in the public record assembled for this review jointly documents actual AI cost, an explicit budget trade-off, a verified workforce action, demonstrated replacement of the displaced work, and realized financial return net of AI total cost. The answer in this search is no. Complete cases = 0. Near-complete cases (four of five links) = 0. Several named organizations are Partial (two or three links). The rest are temporal, opposite-of-substitution, or not substitution cases. Vendor TEI, list prices, promotional OneGov prices, and company estimates such as Klarna’s \$40 million are not scored as realized ROI. Table XXXII and Fig. 22 are the scorecard; Table XXXIII answers the priority sequence question. Full rows live in Enterprise_AI_Longitudinal_Evidence.xlsx.

What the search did find, without upgrading the class. IBM has a 2023 planned back-office hiring pause on expected future automation and later 10-K language that productivity actions ‘allowed for continued investments’; it does not publish an assistant invoice, a completed 7,800-person WARN census, or

TABLE XXX

TWO-LEVEL EVIDENCE MAP. COMPANY CELLS AND ADMINISTRATIVE CELLS ARE NOT INTERCHANGEABLE. IRS/SSA MICRODATA ARE OUT OF SCOPE.

Level	Source	What it can show	What it cannot show
1 Company	10-K / 10-Q / 8-K, earnings, WARN	Named HC, capex, stated freezes, named reductions	Industry hiring rates; other firms' layoffs
1 Company	AI-investment disclosures	Unsplit capex/R&D or promo prices	Buyer AI TCO unless the filing splits it
2 Industry/geo	Census QWI (public use)	Emp, hires, seps, job creation/destruction, payroll by NAICS/geo/size	Named employers; AI treatment
2 Industry/geo	BLS QCEW / CES	Establishment employment and wages	AI spend; company identity in this extract
2 Industry/geo	BLS JOLTS	Openings, hires, quits, layoffs/discharges	Which firms adopted AI; WARN identity
Not used	IRS 941 / SSA employer records	—	Not public company-level datasets; not claimed

TABLE XXXI

TRIANGULATION WITHOUT POOLING UNITS. FULL ROWS: WORKBOOK SHEET ADMIN_TRIANGULATION. NO SENTENCE OF THE FORM "AI CAUSED LAYOFFS" IS SUPPORTED BY THESE ADMINISTRATIVE TOTALS.

Question	Company file	Administrative file	Joint reading
AI \$ then lower HC?	No simple inverse in Panel A; buyer \$-HC n = 0	JOLTS cannot name firms	No spend elasticity
Hiring/openings after 2022?	Tech cuts mostly class 6/7, often pre-Copilot GA	Openings -37%; hires -18%; quits -25%; layoffs +20%	Temporal; confounders first
Information hiring	Overhire/cost language dominates 2022–23	Hires -31% in 2023, then flat	Tech reset, not identified substitution
Are 2025 JOLTS layoffs AI?	Class 1 invoice+FTE still n = 0	21.239 million all-cause	Do not relabel

an independent pre-pause replacement test. Salesforce's 2025 engineer freeze is a stated hiring action with claimed 30% productivity, not an independent gate and not a buyer TCO. Duolingo's contractor non-renewal is verified as a company comment and is paired with a company content-unit series, not an invoice or audited ROI. Bank of America describes attrition mix and a multi-billion technology envelope that is not AI-only. Cisco's 14 August 2024 8-K and call are the clearest U.S. reallocation language ('reallocating versus ... cost savings'; 'shifting hundreds of millions of dollars into AI') beside a verified reduction and up to about \$1 billion of severance-class charges — and they still do not show that production AI already performed the displaced jobs. UiPath (~10% of ~4,200) and Five9 (~4%) 8-K/10-Q text prioritize AI investment while cutting staff; replacement metrics are UNKNOWN. Dropbox's 2023 and 2024 reductions sit beside memos that free capacity for Dash while the 2024 exhibit still calls Dash 'early.' Chegg's 2024–25 8-K/10-K reductions (~760, then a further 248) name student adoption of generative AI as a competitive demand shock and separately report lower content CapEx with more questions asked; that is not a packet that Chegg's GPT-4 absorbed those eliminated roles before the RIF. GSA and IRS are federal: GSA lost a large share of staff in 2025 and in 2026 described a later 'million hours' automation challenge; IRS RAAS lost AI-supporting specialists [59]. OneGov promo prices and the Dillon RCT are not substitution packets. USASpending/SAM.gov extracts in this search did not produce a complete agency TCO-plus-headcount pair.

workforce action? For every row with a claimed or contemporaneous AI workforce action, the public answer is No or UNKNOWN, never Yes (Table XXXIII). The two clearest people-first sequences are international Klarna and U.S. federal GSA. Klarna announced two-thirds of chats handled and a 700 FTE-equivalent in February 2024 after a hiring freeze; in May 2025 the CEO told Bloomberg that cost had been 'too predominant' and quality was lower, and the firm began recruiting humans so customers could always reach a person. That is freeze/automation first → quality problem → human fallback. It is not a complete case: cost and realized ROI remain UNKNOWN, and 700 was never a verified WARN layoff. GSA's 2025 staffing loss preceded the 2026 automation challenge; GAO-cycle reporting found skills and performance analysis after the cuts, service gaps at customer agencies, and hundreds of PBS reduction notices later rescinded. Dropbox cut while Dash was still early. Cisco, UiPath, and Five9 moved people off payroll to fund AI products or R&D without showing that AI already ran those jobs. IBM paused on expected future replacement. IRS removed people who supported AI. The managerial implication is narrow and documented: organizations in this public set have reduced or frozen people without a five-link proof that the system could perform the displaced work at acceptable cost, quality, reliability, and risk. That is not the same sentence as 'AI caused the layoffs,' and it is not a finding that replacement never occurs inside firms. It is a missing-packet result.

Priority question — was replacement proved before the

TABLE XXXII

FIVE-LINK SUBSTITUTION SCORECARD. MODELED TEI AND FORECASTS ARE EXCLUDED FROM REALIZED ROI. FULL ROWS ARE IN THE COMPANION LONGITUDINAL WORKBOOK. COMPLEMENTARY TO PRISMA N = 61.

Company	Class	Actual AI \$	Reallocation	WF action	Replacement	Realized ROI
Complete 5-link case	n = 0	—	—	—	—	—
Near-complete (4/5)	n = 0	—	—	—	—	—
IBM	Partial	UNKNOWN	PARTIAL (10-K)	Planned pause	UNKNOWN	CLAIMED
Salesforce 2025 eng.	Partial	UNKNOWN	UNKNOWN	Stated freeze	CLAIMED	UNKNOWN
Duolingo	Partial	UNKNOWN	UNKNOWN	Contractor ~10%	Company series	UNKNOWN
Cisco 2024 8-K	Partial	UNKNOWN (shift \$)	YES language	YES 8-K	UNKNOWN	UNKNOWN
UiPath / Five9	Partial	UNKNOWN	PARTIAL 8-K	YES 8-K/10-Q	UNKNOWN	UNKNOWN
Dropbox	Partial	UNKNOWN	PARTIAL memo	YES 2023+24	Dash 'early'	UNKNOWN
Chegg	Partial (competition)	UNKNOWN	UNKNOWN	YES ~22%	Content \$ only	No (demand shock)
Klarna (intl)	Partial + reversal	UNKNOWN	UNKNOWN	Freeze/attrition	Claimed then quality fail	No (\$40M estimate)
GSA 2025–26	Partial + sequence	Promo / UNKNOWN	Reverse order	YES ~40%	After the cuts	UNKNOWN
IRS RAAS	Not substitution	UNKNOWN	UNKNOWN	YES -63	Opposite	UNKNOWN

TABLE XXXIII

PRIORITY QUESTION: DID THE ORGANIZATION PROVE REPLACEMENT AT ACCEPTABLE COST, QUALITY, RELIABILITY, AND RISK BEFORE THE WORKFORCE ACTION? NO ROW IS YES. COMPLEMENTARY TO PRISMA N = 61.

Organization	Sequence	Proved before WF action?	Fallback
Klarna	Freeze/automation → claimed 2/3 chats → lower quality → human re-recruiting	No	Yes
GSA	Large 2025 cuts → 2026 automation challenge → GAO skills-after; some RIFs rescinded	No	Partial
Dropbox	RIF to fund Dash → Dash still early in 2024 exhibit	No	UNKNOWN
Cisco / UiPath / Five9	RIF + fund AI products/R&D; no production absorption shown	No	UNKNOWN
IBM	Pause on expected future automation	No	UNKNOWN
IRS RAAS	Cut AI-supporting staff	No (opposite)	Skills gap
Any complete case	None identified	—	—

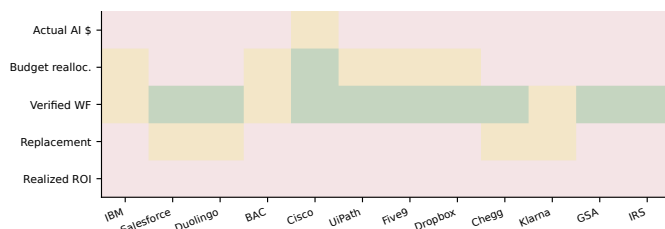


Fig. 22. Completeness grid for named organizations in the targeted search. The realized-ROI row is empty by construction of the acceptance rule. No complete case.

4.27 Actual organizational AI spend across platforms

The complementary question is what organizations actually paid—not what vendors advertised—for the major enterprise AI products, and whether any row completes the chain actual spend → deployment → measured replacement → workforce change → realized financial outcome. Table XXXIV is the master file. List price × guessed seats is never computed. Incomplete rows are kept. Realized ROI is UNKNOWN on every row.

Where actual or award dollars appear, they are almost all public-sector. U.S. Treasury delivery order 2032H325F00060 shows a current award value of \$3,398,315 for up to 3,000 ChatGPT licenses. Separate USASpending extracts in the same compilation show HHS \$252,683.60, the National Gallery of

Art \$95,001.01, and the National Endowment for the Arts \$59,400. Army Intelligence Data Platform delivery orders paid Palantir USG \$333,333.34 for a two-month Artificial Intelligence Platform license (W56KGY25F0083) and \$1,166,666.69 for a seven-month license (W56KGY25F0090). Those are ACTUAL award lines. They are not full TCO (implementation, consulting, training, and security remain UNKNOWN) and they have no paired headcount or realized-ROI file. The 31 July 2025 Army–Palantir Enterprise Agreement is a \$10 billion ten-year ceiling. The Army’s own release states that the figure is a maximum potential, not an obligation. It is coded CEILING / NOT ACTUAL SPEND.

University and state procurement is the next most transparent. A reported University of Colorado ChatGPT Edu order form totals \$2,100,000 for 100,000 end users (January 2026–January 2027). The University of Maine System stated a planned two-year ChatGPT Edu contract of about \$1.39 million covering roughly 25,200 students and 5,600 employees—ESTIMATED / NOT ACTUAL until the executed invoice. Mississippi State’s sole-source notice for Claude Education Enterprise will not exceed \$250,000. The University of Kent awarded OpenAI UK Ltd £3,257,010 excluding VAT (international comparator). California’s Poppy platform is documented as statewide and vendor-agnostic (Claude, Gemini, GPT, Nova) with a 2,800-person pilot, but CDT published no vendor invoice total (existing staff; SITEs discounts). University of Georgia’s published FY26 rates (\$276 ChatGPT Edu / Claude Enterprise;

\$240 Gemini AI Pro; \$276 M365 Copilot) are LIST PRICE ONLY and are not multiplied by guessed seats.

OneGov introductory prices remain committed promotional objects, not commercial TCO: ChatGPT Enterprise \$1 per participating agency-year (6 August 2025); Claude \$1 per agency on the launch release versus later inventory language of \$1 per user (unit conflict left unresolved); Gemini for Government \$0.47 per agency (inventory compilation); Microsoft 365 Copilot \$0 for up to 12 months for qualifying G5 customers, with a GSA-claimed \$3.1 billion potential Microsoft-related saving that is not an invoice; ServiceNow ITSM Pro/Pro Plus (Now Assist included) at up to 70% off list. A 10 September 2026 GSA follow-on moves ChatGPT to 50% off tokens for 27 months with no platform fee—still a discount schedule, not a usage invoice. A \$1.07 million SBA ServiceNow BPA call (Metaphase) is ACTUAL professional-services dollars and is not isolated as Now Assist.

Private-sector named buyers of Agentforce, Now Assist, Copilot Studio, GitHub Copilot, watsonx, Joule, Oracle AI, Adobe Firefly enterprise, Bedrock agents, Vertex / Agentspace, Snowflake Cortex, and Databricks AI remain UNKNOWN as organizational TCO in this search. Vendor logos, supplier PPE (Microsoft additions to PPE \$64.551 billion in FY2025), and Bank of America's unsplit ~\$4 billion technology-initiatives language are not buyer AI ledgers. Dillon's 300-license incentive, the UK 20,000-license evaluation, Virginia Tech's 177 seats, and CSIRO's 300 seats are seats or incentives, not invoices.

Linking spend to workforce and ROI (Table XXXV; bins in Table XXXVI / Fig. 23). Complete chains $n = 0$. Near-complete chains $n = 0$. Organizations with actual award dollars and no realized ROI data include Treasury, HHS, NGA, NEA, Army AIP task orders, the reported CU order, and Kent. Organizations with workforce reductions and no actual AI-spend evidence include Cisco, UiPath, Five9, Dropbox, Chegg, Meta 2023, and Salesforce 2023–24. Organizations where AI programs continued or expanded while specialist human capability fell remain IRS RAAS [59] and GSA (2025 cuts, then a 2026 automation challenge). High advertised numbers that are not actual spend are the \$10 billion Palantir ceiling, the \$3.1 billion claimed Microsoft saving, OneGov nominal dollars, and the \$30 Copilot list price. The public record therefore now contains real procurement amounts for ChatGPT and Palantir AIP in government and higher education. It still does not contain a reconstructed private-sector TCO for named copilots or agents, and it does not contain a realized financial outcome net of those amounts.

4.28 Evidence adequacy for CGEIT-aligned enterprise AI decisions

The testable claim is not that CGEIT is an inadequate framework. Official CGEIT domains, COBIT EDM01–EDM05, NIST AI RMF, [36], and—for federal agencies—OMB [38] already ask for direction and accountability, resource optimization, benefit realization against a baseline, residual-risk optimization, stakeholder engagement, monitoring, and (in the federal high-impact setting) discontinuation when

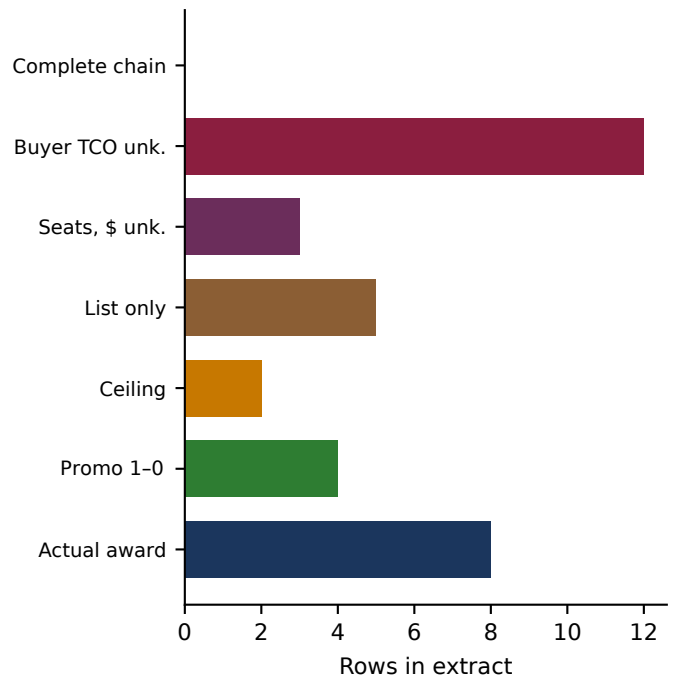


Fig. 23. Public-record spend bins. Actual awards are not complete TCO. Complete spend-to-ROI chains remain $n = 0$.

performance is not appropriate. Those authorities do not fail to request a decision packet. The empirical question is whether enterprise-AI evidence currently supplied to a decision maker is adequate to use that packet for a fully informed, benefits-based, risk-aware authorization of scale-up, budget reallocation, or workforce reduction. Table XXXVII maps each required input; Table XXXVIII is the minimum evidence package; Fig. 24 is the adequacy sketch. This is evidence deficiency, not framework deficiency.

Which decision-critical inputs are missing or insufficient? Against this review: (i) a complete TCO envelope is missing for essentially all private named buyers and is incomplete even where USASpending shows a license award (implementation, consulting, cloud, training, security, maintenance, HITL, and failed-pilot cost remain UNKNOWN); (ii) independently measured financial ROI net of AI TCO was not identified for named enterprise assistants—the Forrester TEI [28], Klarna's \$40 million, and GSA's \$3.1 billion potential Microsoft saving are modeled or claimed; (iii) replacement capability at production quality is not documented before workforce actions (Section 4.26: no public Yes); (iv) reliability, error, escalation, and rollback rates at organizational scale are UNKNOWN (class E breach $n = 0$ verified is not a reliability rate); (v) workforce mechanism plus actual AI \$ remain unpaired (complete substitution cases $n = 0$); (vi) specialist human requirements are under-specified, and in the IRS RAAS case the specialists were removed; (vii) published go/no-go and stop criteria are rare outside study protocols and federal process. What is available is not nothing: Census firm-share use and the 2% AI-related employment-decrease item; local causal productivity (Dillon, Brynjolfsson, Cui) with equally real nulls; independent Copilot UX in UK/CSIRO evaluations; official

TABLE XXXIV

MASTER ORGANIZATIONAL AI-SPEND EXTRACT ACROSS PLATFORMS. FULL COLUMNS ARE IN THE COMPANION SPEND WORKBOOK. COMPLEMENTARY TO PRISMA N = 61.

Organization	Product	Observed \$	Spend type	Seats	ROI / WF
U.S. Treasury OCIO	ChatGPT Ent.	\$3,398,315 award	ACTUAL DO	≤3,000	Both UNKNOWN
HHS / NGA / NEA	ChatGPT-class	\$253k / \$95k / \$59k	ACTUAL orders	UNKNOWN	Both UNKNOWN
U.S. Army AIDP	Palantir AIP	\$333,333; \$1,166,667	ACTUAL DOs	1 license each	Both UNKNOWN
U.S. Army EA	Palantir commercial/AI	\$10B cap	CEILING not paid	N/A	Do not infer
CU (reported form)	ChatGPT Edu	\$2,100,000	Reported order	100,000	Both UNKNOWN
UMaine System	ChatGPT Edu	~\$1.39M / 2 yr	ESTIMATED planned	~30,800 people	UNKNOWN
MSU	Claude Edu Ent.	NTE \$250,000	CEILING notice	>150 prior	UNKNOWN
U. Kent (intl)	OpenAI platform	£3,257,010 ex VAT	ACTUAL award	UNKNOWN	UNKNOWN
OneGov agencies	ChatGPT / Claude / Gemini / Copilot / Now Assist	\$1 / \$1 / \$0.47 / \$0 / % off list	PROMO or LIST	UNKNOWN	HC not in PRs
Private Agentforce/ Studio/Bedrock/ Joule/watsonx buyers	Those platforms	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
M365 Copilot market	M365 Copilot	\$30/user/mo	LIST PRICE ONLY	Do not guess	N/A

TABLE XXXV

SPEND–WORKFORCE–ROI LINKS FOR PUBLIC AWARD AND WORKFORCE ROWS (ABBREVIATED). EVERY COMPLETE FIVE-LINK CHAIN REMAINS $n = 0$. COMPLEMENTARY TO PRISMA $n = 61$.

Organization / row	Spend evidence	Workforce link	Realized ROI
U.S. Treasury ChatGPT award	ACTUAL \$3,398,315 (≤3,000 seats)	UNKNOWN	UNKNOWN
Army Palantir AIP delivery orders	ACTUAL \$333,333 and \$1,166,667	UNKNOWN	UNKNOWN
Cisco / UiPath / Five9 / Dropbox / Chegg	AI \$ UNKNOWN	Verified reductions	UNKNOWN
Complete five-link substitution packet	—	—	$n = 0$

TABLE XXXVI

CLASSIFICATION OF THE SPEND FILE. TABLE XXXV REPEATS UNKNOWN ON WORKFORCE AND ROI LINKS FOR THE AWARD ROWS. COMPLEMENTARY TO PRISMA N = 61.

Bin	n in this search	Examples
Complete \$ → deploy → replacement → WF → realized ROI	0	None
Near-complete (4/5 of that chain)	0	Awards give \$ + product only
Actual award \$; realized ROI UNKNOWN	Several	Treasury; HHS; Army AIP; CU; Kent
Ceiling or promo, not paid TCO	Keep labeled	\$10B Palantir EA; OneGov \$1/\$0; \$3.1B claimed
WF reduction; AI \$ UNKNOWN	Several	Cisco; UiPath; Five9; Dropbox; Chegg; Meta 2023
AI path continues; specialist capability down	Documented	IRS RAAS; GSA 2025–26
Seats known; price unknown	Keep ESTIMATED	UK 20,000; VT 177; CSIRO 300 — not × \$30

risk advisories (NIST, CISA); some actual public ChatGPT and Palantir AIP awards; 8-K workforce actions without AI invoices. Those objects do not combine into a CGEIT-style high-stakes packet.

Can a decision maker reliably authorize AI scale-up, budget reallocation, or workforce reduction when actual TCO, realized ROI, replacement capability, reliability, and workforce impact are unknown? Not as a benefits-based and risk-aware CGEIT/COBIT-style authorization. They can still decide—

under uncertainty, on a vendor roadmap, or on a promotional price. That is an authorized act with an inadequate packet, not a demonstration that the governance framework is empty. EDM02 cannot compare realized benefit with incremental resource use if both sides are UNKNOWN. EDM03 cannot optimize residual risk if production error and rollback are UNKNOWN. EDM04 cannot optimize people versus technology if licenses are list price and specialist labor is uncounted. EDM05 cannot show stakeholders a measured customer or employee effect for

TABLE XXXVII
DECISION-CRITICAL INPUTS. FULL ROWS AND QUALITY LABELS ARE IN THE COMPANION CGEIT WORKBOOK. FRAMEWORK ASKS; PACKET OFTEN DOES NOT ANSWER. COMPLEMENTARY TO PRISMA $n = 61$.

Required evidence	CGEIT/COBIT/NIST/ GAO	Available in this review	Missing	Decision risk
Owner, objective, pre-AI baseline	Dom. 1/3; EDM01–02; GAO P1/P3	Partial (RCTs; speech)	Most commercial baselines	No later benefits test
Complete TCO envelope	Dom. 2; EDM04	Some public awards; else list/promo/UNKNOWN	Private TCO; add-on lines	Wrong cash number
Independent productivity + \$ ROI net of TCO	Tasks 31–33; EDM02; MEASURE	Local RCTs; \$ ROI $n = 0$ independent	Named-agent org ROI	Scale-up on TEI
Replacement + reliability + fallback before WF	Dom. 4; EDM03; NIST; CISA	Insufficient (Klarna reversal; no public Yes)	Pre-action production packet	People-first failure
Specialist HR / knowledge retention	EDM04; [59]	IRS opposite case; Net FTE unknown	Production staffing model	Cut the people AI needs
Workforce mechanism + paired AI \$	Dom. 2–3; [14] unit rule	Cuts yes; complete 5-link $n = 0$	Invoice + mechanism	False “AI-caused” label
Residual risk, rollback, vendor exit	Dom. 4; EDM03; NIST MANAGE	Lab + advisories; org rates UNKNOWN	Production incident/ rollback	Agents as writing tools
Go/no-go, scale-up, stop criteria	EDM01; GOVERN; [38] federal	Study/federal process only	Named-platform stop rules	Renewal by inertia

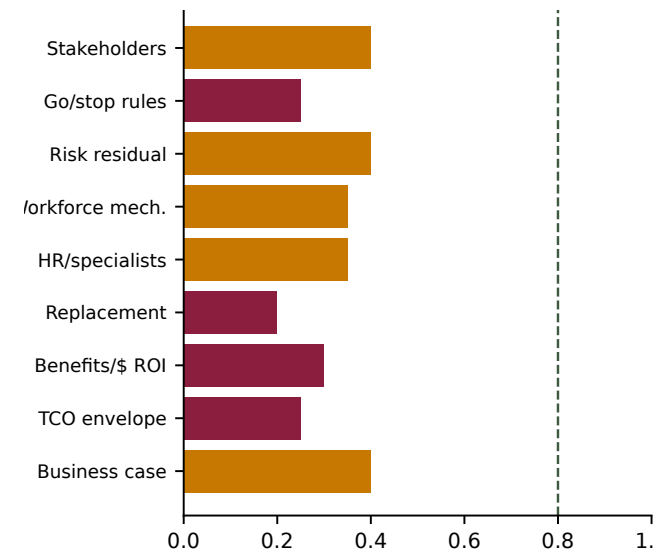


Fig. 24. Public-evidence adequacy by decision area. The dashed marker is a visual aid, not an official ISACA threshold. The figure indicates evidence deficiency, not CGEIT failure.

named agent platforms. Domain 1 cannot maintain an off-ramp that was never written down. The people-first sequences in Section 4.26 are the operational illustration: freeze or cut first, prove replacement later or not at all.

Supported conclusion (revised to match the evidence, not the stronger slogan). The public record supports this sentence and not a stronger one: Incomplete evidence constrains, rather than eliminates, the ability of CGEIT-aligned decision makers to make fully informed, benefits-based, risk-aware enterprise AI decisions. ‘Eliminates’ would over-claim: pilots, promo-year licenses, and local RCTs can still be authorized with partial information. ‘Fully informed, benefits-based, risk-aware’ is the phrase the frameworks imply for irreversible scale-up, budget shifts, and headcount actions—and that standard is not met when TCO, realized ROI, replacement, reliability, and workforce mechanism are UNKNOWN. The claim is not that CGEIT fails. The claim is that the evidence supplied to the decision maker is inadequate relative to what those authorities already ask.

5. DISCUSSION

Three results should constrain interpretation. First, nationally representative U.S. firm data still describe a minority-adoption technology used primarily to augment tasks, with rare self-reported employment decreases [14]. Second, the best causal studies show genuine but local productivity effects and equally genuine nulls and quality reversals [16], [18], [19], [49]. Third, the loudest workforce-reduction statistics in the 2026 analyst stream are organization-share statistics among large firms already pursuing autonomy, and they do not demonstrate that cuts produce returns [24].

Temporally, consumer generative AI (November 2022) and a technology-sector layoff wave preceded Microsoft 365 Copilot general availability (November 2023). Paid enterprise assistants and, later, vendor agent platforms then became purchasable while U.S. firm adoption remained a minority of firms through early 2026. Junior software vacancies and early-career employment in AI-exposed occupations weakened over 2023–2025 in designs that control for some firm shocks but cannot name Copilot or Agentforce as the treatment. Labor-market deterioration therefore both preceded and coincided with enterprise commercialization; coincidence is not contribution, and contribution is not named-product causation. Section 4.13.1 applies that rule case by case: no class A (direct AI-caused layoff) was assigned; IBM and Gartner items are class B; IRS is class D; 2022–23 tech cuts are class E/D; paired invoice-and-layoff ledgers remain $n = 0$. Where the source does not establish mechanism, this review reports temporal association only; causal attribution not established. Agents amplify the same problem: most independently measured benefits are still assistant-class text and suggestions, while the organizational risk sits in unproven writes to enterprise systems (Section 2.2).

These results are jointly incompatible with a simple substitution narrative and with a simple transformation narrative. They are compatible with a jagged, management-mediated technology whose organizational value depends on task selection, human review, enablement, and measurement. Weak governance, poor resource allocation (including loss of AI-skilled staff), absent benefits measurement, and thin risk controls appear in unsuccessful or mixed cases (IRS skills gaps; Klarna quality; predicted agentic cancellations; NANDA

TABLE XXXVIII

MINIMUM EVIDENCE PACKAGE FOR CGEIT-ALIGNED ENTERPRISE AI DECISIONS. THE PACKAGE REPRESENTS THE EVIDENCE THAT CGEIT/COBIT/NIST/GAO-STYLE GOVERNANCE WOULD REQUIRE FOR INFORMED DECISIONS; AVAILABILITY INDICATES WHAT THIS REVIEW IDENTIFIED. THIS IS AN EVIDENCE-ADEQUACY SCORECARD, NOT AN OFFICIAL ISACA FORM OR A FINDING THAT CGEIT FAILS. COMPLEMENTARY TO PRISMA $n = 61$.

Decision Area	Required Evidence	Available?	Evidence Strength	Missing Information	Decision Consequence
Business case	Objective; baseline; owner; alternatives	Partial	Independent in few RCTs; else UNKNOWN	Owner/baseline for most products	No benefits comparison later
Total cost of ownership	Negotiated licenses + impl. + cloud + consulting + training + security + HITL + rework	No (complete envelope)	Official awards (some public); list/promo else	Private TCO; add-on lines	Cannot run EDM04 on cash
Benefits realization	Independent productivity/quality; realized \$ ROI net of TCO; modeled-vs-measured	Partial / No for \$ ROI	Tier A/B local; TEI/ claims excluded as realized	Audited ROI; agent org outcomes	Benefits-based scale-up unsupported
Replacement capability	Production coverage, accuracy, reliability, escalation, HITL, fallback before any WF action	Insufficient	Vendor volume; intl reversal; RCT nulls	Pre-action production test	Not risk-aware headcount authority
Human-resource requirements	Builders, IAM, reviewers, monitors, maintainers, reskilling, knowledge retention	Partial / opposite	GAO IRS; UK friction; Net FTE insufficient	Named-org staffing model	Licenses without the operators
Workforce consequences	Mechanism-coded action; paired AI \$ and HC; expertise check	Partial (cut $\neq$ mechanism)	SEC/GAO for cuts; 5-link $n = 0$	Invoice + replacement + ROI	Cannot say “AI-caused”
Risk	Cyber, privacy, error, ops, vendor, compliance, continuity, rollback	Partial	Official NIST/CISA/ GAO; org rates UNKNOWN	Residual-risk + rollback at scale	Agents approved on writing-tool risk
Governance and accountability	Owner; approval; audit trail; monitoring; human authorization; go/no-go; stop	Partial	Federal process; private gate = paper proposal	Named-platform stop/ scale rules	No off-ramp
Stakeholder impact	Employee, customer, unit, public-user, capability effects	Partial	Independent Copilot UX; Klarna/GSA negatives	Agent-platform independent packs	EDM05 has little to show
Evidence quality	Do not pool independent / official / audited / self-report / vendor / modeled / UNKNOWN	Coded here; rare in source packets	This review’s protocol	TEI presented as realized ROI	False certainty

production stall). We report that pattern because it is repeated across independent source classes; we do not claim a structural identification of governance as the cause of ROI.

A fourth constraint is the evidence-maturity gap between Category 1 and Category 2 (Section 4.22). Independent organizational measurement is currently thicker for systems that assist a human than for systems that act inside the organization. That is not a finding that agents fail, and it is not a finding that assistants pay for themselves. It is a finding about what can be claimed. Transferring Copilot minute-savings or customer-service issues-per-hour into an Agentforce or Copilot Studio business case would merge objects the protocol keeps apart. The same gap explains why governance and security language in official advisories is more demanding once tools can write: autonomy raises the cost of being wrong faster than the independent outcome literature has grown.

Section 4.23 shows a second, reporting gap. Product-level conclusions that would be straightforward if scale, use, TCO, productivity, quality, ROI, reliability, HITL, security, and workforce were published together cannot be drawn from the included public record. Assistant studies more often disclose an experimental n and an operational metric; they still omit invoices and realized financial return. Named agent platforms more often disclose a launch date or a logo. Surveys cannot close the gap: they use different denominators and almost never observe actual cost and actual headcount on the same named system. The correct inference is limited claim strength, not a hidden true ROI.

That fragmentation is the governance result. Sections 4.4 and 3.4 keep workforce attribution split (substitution, compression, non-backfill, budget reallocation, contribution, AI-present-without-mechanism, temporal only); most large 2022–23 cuts remain in the last two bins, and the cases that survive alternatives are mostly hiring pauses, contractor non-renewal,

and attrition-with-output statements—not WARN-scale “AI layoffs.” Section 4.24 could not estimate a buyer-spend DiD ($n = 0$ complete pairs); absence of an elasticity is missing data, not a zero effect. The same subsection shows why supplier PPE and buyer TCO cannot be stacked. Section 4.25 shows national JOLTS cooling after 2022 without identifying named-firm AI displacement. Section 4.26 found complete five-link substitution cases $n = 0$ and several people-first sequences (Klarna; GSA; Dropbox/Dash; IRS). Section 4.27 found actual public awards (Treasury ChatGPT; Army Palantir AIP; university ChatGPT Edu) with UNKNOWN realized ROI, and private platform TCO still UNKNOWN. List price, promotional ceilings, and claimed savings remain distinct objects (Section 3.4). Together these are not a second results section; they are why a CGEIT/COBIT-aligned board cannot treat the public packet as adequate for irreversible scale-up or headcount actions while still being able to authorize a measured pilot.

For organizations, the practical implication is narrower than vendor roadmaps and broader than a single RCT. License provision can reduce some email time without changing measured task quantity [49]. Licensed public-sector employees often report high satisfaction with Microsoft 365 Copilot and want to keep the tool, yet independent evaluations do not show that self-reported minutes become independently verified productivity or financial return (GDS, 2025; Department for Business and Trade, 2025). Agent programs that treat headcount reduction as the value thesis lack supporting ROI evidence in the included analyst set. Security and identity work—least privilege, dedicated agent identities, human approval of high-impact actions—is an operational requirement in official advisories (CISA and partners, 2026; NIST NCCoE, 2026), not an optional maturity add-on. A working prototype is not, on this evidence, a successful enterprise deployment.

For researchers, the contribution is a conservative identifi-

cation strategy rather than a new point estimate. The review shows that mixing Census firm shares, McKinsey respondent shares, and Gartner organization-share layoff items produces a spurious “AI adoption and jobs” fact. It also shows that vendor-affiliated Copilot satisfaction numbers cannot be pooled with independent departmental evaluations. Future designs listed in Section 8 follow directly from these gaps: U.S. controlled public-sector evaluations, invoice-linked TCO panels, matched employer–employee data, and pre-registered production-agent safety endpoints.

For enterprise AI governance, CGEIT remains a lens. Access is easier to observe than benefits realization or residual-risk optimization. That descriptive imbalance is the closest the included set comes to an adoption-ahead-of-governance claim, and it is not a certified finding about any named organization. Boards can still ask the four questions in Section 6.2 without claiming compliance. Section 4.28 states the adequacy result without saying CGEIT fails: incomplete evidence constrains, rather than eliminates, fully informed, benefits-based, risk-aware enterprise AI decisions. Scale-up, budget reallocation, and workforce reduction cannot be reliably authorized on that standard while TCO, realized ROI, replacement, reliability, and workforce mechanism are UNKNOWN.

International evidence was used as specified. Independent UK departmental Copilot evaluations are the largest public user-experience evidence on the product and are labeled throughout as United Kingdom evidence. BCG, CSIRO, Klarna, Kenya, Singapore, and the mostly Philippine customer-service workforce are similarly labeled. They enrich mechanism evidence; they do not substitute for U.S. official adoption and workforce shares.

5.1 Independent user-experience synthesis answers

1. Are enterprise AI users actually satisfied? For licensed Microsoft 365 Copilot employees in independent public-sector evaluations, satisfaction and keep-the-tool items are generally high (GDS 82% would keep; DWP 95% satisfied; DBT 72%; HMRC 7.1/10). Findings are mixed by task. Insufficient comparable independent evidence exists for other named enterprise assistants. 2. Are agent builders satisfied? Insufficient independent empirical evidence was identified. 3. Is using an assistant easier than building an agent? Using Copilot is documented as often easy to start for Office-familiar employees; building agents is an evidence gap. That is not a comparative ranking. 4. Can nontechnical users create agents without substantial IT support? Insufficient independent empirical evidence was identified. 5. Can they move from prototype to production? Independent sources document LEVEL 2 departmental assistant use, not LEVEL 3–4 agent progression. 6. Where does complexity increase most? Where evidence exists, friction is documented at permissions, security/classification, uneven app/integration quality, and (for some tasks) output quality—not at first-login for many employees. Later builder stages are unmeasured. 7. Largest independently documented frictions for employees: configuration/integration (unused Excel; poor slides/scheduling), security/privacy non-use, prompting/learning, and the gap between minutes saved

and productivity. Training is MIXED (self-led often used). Reliability and cost are not independently established as the dominant UX complaint. 8. Does high satisfaction translate into sustained adoption? Short-trial continued-use interest is high (GDS ~80% active; DWP 77% extremely interested in extension; HMRC 64% increased use; 61% would miss the licence). Post-pilot year-long independent persistence is not measured. 9. Does easy demo creation translate into enterprise agent deployment? Insufficient independent empirical evidence was identified. 10. Gaps: citizen-developer Copilot Studio studies; professional-developer integration UX; IT/security administrator satisfaction; LEVEL 3–4 agent deployments; comparable independent surveys of Salesforce, Google, AWS, OpenAI, Anthropic, and ServiceNow builders.

6. IMPLICATIONS

6.1 For executives and public managers

Treat list prices, promotional federal prices, and commissioned TEI composites as different numbers. Do not use organization-share layoff statistics as employee-share statistics. If the value thesis is headcount reduction, the included analyst evidence does not show that this thesis yields higher ROI. If the value thesis is augmentation, the included causal evidence is more supportive—but still requires a plan for how saved minutes become output, quality, or service, which the UK evaluation could not observe. Agent deployments should assume prompt-injection and permission-inheritance risk and should specify human authorization, logging, and rollback before autonomous actions touch customers, money, or regulated data. Treat agents as action-taking systems: if the tool can change a ticket, a record, or an entitlement, the decision is a risk decision, not a writing tool purchase (Sections 2.2 and 4.22; Fig. 25). Distinguish “AI that assists a human” from “AI that acts inside the organization.” Do not use Category 1 productivity or satisfaction results as Category 2 proof. For agents, inventory autonomy, tool access, permissions, record-change and workflow-trigger rights, approval points, rollback, monitoring, testing, identity, and independent production evidence before scale-up. Decision rule (proposal of this review): the more consequential the agent’s action authority, the stronger the evidence, controls, monitoring, and human authorization required before production scale-up. A workforce action may not be labeled “AI-caused” unless a documented mechanism is in the packet (Section 4.4.2 classes 1–5). Timing plus an AI slide is class 6 or 7. Before a freeze, non-backfill, or layoff is attributed to AI, ask the Section 4.26 question: was replacement proved at acceptable cost, quality, reliability, and risk first? In the public cases assembled here, the answer was not Yes.

6.2 For boards and assurance functions

CGEIT-oriented oversight can ask four questions that this literature actually illuminates: Who owns the AI decision and the acceptable-use policy? What is the full resource envelope (licenses, credits, integration, training, specialist staff)? Which benefits are expected, claimed, measured, and independently verified? What is the residual risk and the fallback when the model is wrong? These questions do not require claiming

CGEIT certification. A director or board should require a staged packet before first investment, before a pilot, before production, before scale-up, before budget expansion, and before workforce restructuring (Table XXXVIII). Section 4.15.1 translates the same principles into an evidence-gated sequence—initial investment, pilot, production or scale-up, renewal or budget expansion, then workforce restructuring—and states explicitly that the gate itself is a proposal of this review. No examined official source imposes a 3-, 6-, or 12-month waiting period. Boards should therefore ask whether realized benefits, total cost, residual risk, and a documented workforce-attribution mechanism are in the packet before authorizing enterprise-wide licensing or irreversible headcount actions, and should refuse the label “AI-caused layoff” when only temporal association or budget displacement is shown. Table XVIII is the source-by-source record; this subsection does not repeat it. Section 4.13.1 is the longitudinal test of those labels; Section 6.4 is the one-page managerial checklist (a proposal, not a CGEIT certificate). Table XXXVIII is the minimum evidence package those questions imply; availability is an evidence-deficiency scorecard, not a framework failure.

6.3 For workers and labor-market institutions

Entry-level hiring pauses reported by a minority of CHROs, attrition-based service-center reductions, and high BYOAI use are early labor-market signals. They are not, in U.S. official data through early 2026, a measured mass displacement wave caused by Copilot-class tools. Academic series on junior software vacancies and early-career employment in AI-exposed occupations [61], [62] are AI-contributing evidence of uneven entry-level demand, not proof that a named assistant caused a given layoff. Reskilling claims should be evaluated with the same skepticism as ROI claims: hours, participation, and subsequent job outcomes are rarely reported. National JOLTS cooling after 2022 is an industry/geography administrative fact; it is not a substitute for a named-employer mechanism (Section 4.25).

6.4 One-page executive decision guideline (proposal of this review)

Fig. 25 is a one-page checklist for managers and boards. It is based primarily on the four CGEIT domains—governance of enterprise IT, IT resources, benefits realization, and risk optimization. COBIT EDM01–05, NIST AI RMF, [36], the CISA agentic advisory [33], and OMB [38] (federal agencies only) are supporting authorities. The decision gate and the checklist are a framework proposed by this review. They are not an official CGEIT requirement, not a certification, and not a 3-/6-/12-month clock.

For every major decision, ask five questions in business language. What evidence do I need? What do I check? What would justify scale-up? What would justify stopping? What evidence is required before reducing headcount? First spend needs an owner, a pre-AI baseline, a total cost, and a clear call on assistant versus agent. More seats need a realized metric, not a modeled ROI. Write-backs need an agent identity, a log, a rollback, and a person on high-impact actions. Growing

the AI budget while cutting other work is a resource choice: write down what you de-funded. Freezing hiring or laying people off and calling it AI requires a written class A, B, C, or D. If you cannot write it, it is class E. The default sentence remains: Temporal association only; causal attribution not established. If the system can act, apply the Section 6.1 rule: more action authority requires stronger evidence and stronger human authorization before scale-up. Do not mark a freeze, attrition program, or layoff “AI-caused” without a written mechanism (substitution, compression, non-backfill, or budget reallocation). If the mechanism is missing, the label is temporal association only.

7. LIMITATIONS

Several constraints are inherent to the search environment and are disclosed rather than repaired after the fact. Scopus and Web of Science were not institutionally available; identification therefore relied on institutional hand-search, targeted bibliographic and web queries, Crossref DOI verification, and citation snowballing. That limitation can omit studies that those databases would have surfaced. Some journal version-of-record PDFs and all Gartner subscription reports were inaccessible; we used author-accepted or working-paper PDFs, official abstracts, and newsroom text only, and we did not circumvent paywalls. The official PRISMA identification and exclusion counts are reconciled process tallies (Supplementary S2). They are internally consistent with the 61-record included set but are not a 471-row unique-ID export.

Evidence quality is uneven by design of the literature, not only of this review. Vendor materials may omit failures; they remain in the corpus for labeled non-UX questions and are excluded from the primary UX synthesis. Self-report time savings can diverge from telemetry. Independent Copilot UX studies are concentrated in UK government departments and one Australian research agency; they generally lack randomized assignment (HMRC allocated licences randomly, but outcomes remain survey-based). Several Tier A Copilot and GitHub studies have Microsoft-affiliated authors and are therefore VENDOR-AFFILIATED for UX even when they are usable for other research questions. The absence of independent Copilot Studio builder studies is a finding, not a warrant to import vendor ease-of-use claims. National BTOS question wording changed in November 2025, so time series of “AI use” are not fully comparable across vintages. The organization-level spend–workforce file is sparse because public sources rarely publish paired invoices and verified layoffs. The complexity deep-dive still lacks independent named-organization builder studies for ServiceNow, Vertex, Bedrock, watsonx, Joule, and Oracle agents. No meta-analysis was performed because constructs, samples, and bias structures were incommensurate. The review was not registered in PROSPERO. The 2022–2026 commercialization, attribution-chain, and agentic-task inventories are targeted updates complementary to the 61-record PRISMA set; they do not change identification $n = 471$. Private-company assistant invoices are rarely disclosed; SEC capex is vendor infrastructure. Job-posting series and payroll series are not the same as layoffs. Macroeconomic

One-page executive decision guideline (proposal of this review)

Based primarily on the four CGEIT domains. Not an official CGEIT, COBIT, NIST, GAO, or OMB requirement.

Decision	CGEIT domain focus	Evidence you need	Scale-up only if...	Stop if...	Headcount cut?
First spend / pilot	Governance; Resources	Owner, pre-AI baseline, TCO, and assistant-versus-agent classification	Baseline and ownership are established; use case is clearly defined	No owner, no baseline, or only a demo	No
More seats / production	Benefits; Resources	Realized metric versus baseline, adoption, errors, quality, and actual TCO	Metric improves without unacceptable quality loss	Null metric, quality deterioration, or only modeled ROI	No. Minutes saved $\neq$ FTE
Allow write-backs / agentic actions	Risk; Governance	Systems touched, agent identity, permissions, logs, rollback, and human approval	Workflow is bounded, auditable, reversible, and human-authorized where needed	Shared credentials, no audit trail, excessive privileges, or no rollback	No. Autonomy $\neq$ labor savings
Grow AI budget / cut other programs	Resources; Benefits	Actual invoices, incremental TCO, realized benefits, and what is being de-funded	Realized benefit exceeds incremental cost and the trade-off is documented	Only list price, promotional pricing, modeled TEI, or projected savings	Only as an explicit budget-reallocation decision; not proof that AI replaced work
Freeze hiring / lay off as "AI"	All four domains	Production evidence, replacement capability, quality, reliability, fallback, actual cost, workforce mechanism, and attribution class	Displaced work is demonstrated in production at acceptable cost, quality, reliability, and risk	Evidence is only timing, surveys, pilots, or examples from other organizations	Only with a documented mechanism; otherwise temporal association only

Check every time: Distinguish text assistance from actions that change enterprise systems. Demo, pilot, production, and independently evaluated production are different. Do not mix list price, promotional pricing, modeled TEI, vendor capex, and actual invoices. Organization-share surveys are not percentages of employees.

Supporting authorities: COBIT EDM01–EDM05; NIST AI RMF; GAO AI Accountability Framework; CISA agentic-AI guidance; OMB M-25-21 for federal agencies. WARN is a notice law, not an AI-readiness test.

Fig. 25. One-page executive decision guideline. Proposal of this review, aligned to the four CGEIT domains. Not an official ISACA requirement.

confounding cannot be fully removed. WARN and 8-K layoff universes were not exhaustively coded. Public QWI stops short of named employers and, in the R2026Q1 vintage used as the reference window, national files lag at about 2025Q1. A complete QCEW NAICS-quarter extract was not rebuilt for an invalid AI-treatment model. JOLTS annual NSA tables are the official industry/geography flow file in Section 4.25; monthly seasonally adjusted vintages differ slightly and June 2026 openings are preliminary. IRS and SSA employer tax records were not obtained and are not claimed. The five-link substitution search used public EDGAR, earnings, GAO, and GSA/OneGov records; it did not obtain confidential invoices or a complete national WARN-as-AI census. Absence of a complete case is a public-record result, not proof that no firm internally holds such a packet.

8. FUTURE RESEARCH

Priority designs include: (1) U.S. public-sector Copilot and ChatGPT Enterprise evaluations with control groups, telemetry, and time-use diaries that track reallocation; (2) independently audited TCO and ROI panels linking invoices to operational metrics; (3) matched employer–employee data that can test whether AI expenditure precedes employment change after conditioning on demand shocks, including confidential LEHD–BTOS or LEHD–SEC linkages that this public review cannot construct; (4) production evaluations of named agent platforms with pre-registered safety and quality endpoints; (5) incident registries for prompt injection and incorrect autonomous actions; (6) entry-level hiring studies that measure job counts, not CHRO anecdotes; and (7) governance field experiments that randomly assign measurement and oversight practices, rather than inferring CGEIT effects from observational success stories.

9. CONCLUSION

From January 2022 through September 2026, the public record on enterprise AI assistants and agents is more consistent with local, task-specific augmentation and incomplete benefits realization than with economy-wide substitution or independently verified financial return. Assistive systems have the thicker independent outcome literature; action-taking agents have the thinner packet and the higher operational stakes. The evidence does not support a predetermined story that these systems primarily destroy jobs, nor that they reliably pay for themselves. Value, risk, and workforce effects are jointly determined by task choice, human authority, measurement, and governance—and those complements are still unevenly in place. The distinctive finding is evidence fragmentation: actual TCO, realized ROI, replacement capability, reliability, workforce mechanism, and residual risk are rarely observed together (Sections 4.22–4.28). Incomplete evidence constrains, rather than eliminates, the ability of CGEIT-aligned decision makers to make fully informed, benefits-based, risk-aware enterprise AI decisions. That is evidence deficiency, not framework failure. Until the minimum evidence package is filled, a board can still authorize a measured pilot; it cannot truthfully describe irreversible scale-up, budget displacement, or headcount reduction as a fully informed CGEIT-style authorization.

EVIDENCE CLASSIFICATION AND GOVERNANCE INTERPRETATION

This compilation preserves the source manuscript's distinctions among independently measured evidence, official government evidence, analyst evidence, vendor-affiliated evidence, modeled ROI, claimed ROI, realized ROI, actual expenditure, list price, estimates, and UNKNOWN. CGEIT

is used solely as an interpretive governance lens and does not constitute certification or endorsement of any organization. The distinction between framework deficiency and evidence deficiency is also retained: established governance and risk-management frameworks such as CGEIT, COBIT, NIST, and GAO emphasize documented evidence, risk assessment, governance controls, and decision support, while the public enterprise-AI evidence base often does not provide sufficient information to fully support those decision requirements.

ACKNOWLEDGMENT

The authors gratefully acknowledge the University of Houston, University Information Technology (UIT), particularly Technology Services and Support, for providing professional and institutional context relevant to this study. The authors also thank colleagues whose discussions and practical insights helped inform the research questions and interpretation of enterprise AI governance, productivity, and investment considerations.

REFERENCES

- [1] OpenAI, "Chatgpt: Optimizing language models for dialogue," 2022.
- [2] —, "Introducing chatgpt enterprise," 2023. [Online]. Available: <https://openai.com/index/introducing-chatgpt-enterprise/>
- [3] Microsoft, "Microsoft 365 copilot is generally available. microsoft tech community," 2023.
- [4] GitHub, "Github copilot for business is now available," 2023.
- [5] Microsoft, "Introducing dynamics 365 copilot," 2023.
- [6] ServiceNow. Product announcements for Now Assist, "–24) and ai agents (knowledge 2025). vendor availability only," 2023.
- [7] Microsoft, "Introducing microsoft copilot studio (ignite)," 2023.
- [8] Salesforce, "Salesforce's agentforce is here (investor news)," 2024.
- [9] Google, "Gemini at work 2025: Introducing gemini enterprise," 2025. [Online]. Available: <https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/gemini-at-work-2025/>
- [10] ISACA, "Cgeit examination content outline (effective 2020)," 2020. [Online]. Available: <https://www.isaca.org/credentialing/cgeit/cgeit-exam-content-outline>
- [11] —, "Cobit 2019 framework: Governance and management objectives (edm01–edm05). schaumburg, il: Isaca," 2018.
- [12] A. Bick, A. Blandin, and D. J. Deming, "The rapid adoption of generative ai (nber working paper no. 32966)," Tech. Rep., 2025. [Online]. Available: <https://doi.org/10.3386/w32966>
- [13] K. Bonney *et al.*, "Tracking firm use of ai in real time (ces 24-16r / nber working paper no. 32319)," Tech. Rep., 2024. [Online]. Available: <https://doi.org/10.3386/w32319>
- [14] K. Bonney, C. L. Breaux, E. Dinlersoz, L. S. Foster, J. C. Haltiwanger, and A. A. Pande, "The microstructure of ai diffusion (ces-wp-26-25). u.s. census bureau," 2026.
- [15] K. McElheran, J. F. Li, E. Brynjolfsson, Z. Kroff, E. Dinlersoz, L. S. Foster, and N. Zolas, "Ai adoption in america: Who, what, and where. journal of economics & management strategy, 33, 375–415," *Journal of Economics and Management Strategy*, 2024. [Online]. Available: <https://doi.org/10.1111/jems.12576>
- [16] E. Brynjolfsson, D. Li, and L. Raymond, "Generative ai at work. the quarterly journal of economics, 140(2), 889–942," *The Quarterly Journal of Economics*, 2025. [Online]. Available: <https://doi.org/10.1093/qje/qjae044>
- [17] S. Noy and W. Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence. science, 381(6654), 187–192," *Science*, 2023. [Online]. Available: <https://doi.org/10.1126/science.adh2586>
- [18] K. Z. Cui, M. Demirer, S. Jaffe, L. Musolff, S. Peng, and T. Salz, "The effects of generative ai on high-skilled work: Evidence from three field experiments with software developers. management science," *Management Science*, 2026. [Online]. Available: <https://doi.org/10.1287/mnsc.2025.00535>
- [19] F. Dell'Acqua, E. McFowland, E. III, Mollick, H. Lifshitz-Assaf, K. C. Kellogg, S. Rajendran, L. Kraye, F. Candelon, and K. R. Lakhani, "Navigating the jagged technological frontier. organization science," *Organization Science*, 2026. [Online]. Available: <https://doi.org/10.1287/orsc.2025.21838>
- [20] N. G. Otis, R. Clarke, S. Delecourt, D. Holtz, and R. Koning, "The uneven impact of generative artificial intelligence on entrepreneurial performance. management science," *Management Science*, 2026. [Online]. Available: <https://doi.org/10.1287/mnsc.2024.06909>
- [21] Reuters, "Ibm to pause hiring in plan to replace 7,800 jobs with ai, bloomberg reports," 2023.
- [22] "Duolingo spokesperson comment to bloomberg (8 jan. 2024) on non-renewal of about 10% of contractors, in part attributable to ai; duolingo company strategy blog on 2024 content-unit volume," 2024.
- [23] Gartner, "Gartner survey finds ai automation is reducing some entry level hiring at nearly one-quarter of organizations," 2026.
- [24] —, "Gartner says autonomous business and ai layoffs may create budget room, but do not deliver returns," 2026.
- [25] U.S. Bureau of Labor Statistics, "Job openings and labor turnover survey: annual tables 15, 17, 21, and 23, 2026 m01 results (not seasonally adjusted)," 2026. [Online]. Available: <https://www.bls.gov/jlt/>
- [26] —, "Job openings and labor turnover survey home page (june 2026 preliminary openings 7,359,000)," 2026. [Online]. Available: <https://www.bls.gov/jlt/>
- [27] Microsoft, "Microsoft 365 copilot plans and pricing—ai for enterprise," 2026. [Online]. Available: <https://www.microsoft.com/en-us/microsoft-365-copilot/pricing/enterprise>
- [28] Forrester Consulting, "The total economic impact of microsoft 365 copilot (commissioned by microsoft)," 2025.
- [29] Tabassi, E, "Artificial intelligence risk management framework (ai rmf 1.0) (nist ai 100-1)," Tech. Rep., 2023. [Online]. Available: <https://doi.org/10.6028/NIST.AI.100-1>
- [30] C. Autio, R. Schwartz, J. Dunietz, S. Jain, M. Stanley, E. Tabassi, P. Hall, and K. Roberts, "Artificial intelligence risk management framework: Generative artificial intelligence profile (nist ai 600-1)," Tech. Rep., 2024. [Online]. Available: <https://doi.org/10.6028/NIST.AI.600-1>
- [31] National Institute of Standards and Technology, "Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (nist ai 100-2e2025)," Tech. Rep., 2025. [Online]. Available: <https://doi.org/10.6028/NIST.AI.100-2e2025>
- [32] Center for AI Standards and Innovation, "Strengthening ai agent hijacking evaluations. national institute of standards and technology," 2025. [Online]. Available: <https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations>
- [33] CISA, ASD's ACSC, and partners, "Careful adoption of agentic ai services," 2026. [Online]. Available: <https://www.cisa.gov/resources-tools/resources/careful-adoption-agentic-ai-services>
- [34] National Vulnerability Database, "Cve-2025-32711 (echoleak / m365 copilot)," 2025.
- [35] "Isaca. certification exam candidate guide (appendix d: Cgeit supporting tasks)," 2026. [Online]. Available: <https://www.isaca.org/credentialing/-/media/fa494652c5f149289af38cef18328650.ashx>
- [36] U.S. Government Accountability Office, "Artificial intelligence: An accountability framework for federal agencies and other entities (gao-21-519sp)," Tech. Rep., 2021. [Online]. Available: <https://www.gao.gov/products/gao-21-519sp>
- [37] —, "Artificial intelligence: Generative ai use and management at federal agencies (gao-25-107653)," Tech. Rep., 2025. [Online]. Available: <https://www.gao.gov/products/gao-25-107653>
- [38] U.S. Office of Management and Budget, "M-25-21: Accelerating federal use of ai through innovation, governance, and public trust," 2025. [Online]. Available: <https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf>
- [39] Anthropic. Claude Team and Claude Enterprise product pages, "–25). vendor availability only," 2024.
- [40] GitHub, "Github copilot enterprise is now generally available," 2024.
- [41] E. DeBenedetti, B. M. Zhang, J., L. Beurer-Kellner, M. Fischer, and F. Tramèr, "Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. neurips 2024 datasets and benchmarks," 2024.
- [42] U.S. District Court for the Southern District of New York, "Mata v. avianca, inc.: Opinion and order on sanctions," 2023, no. 22-cv-1461 (PKC), June 22, 2023. [Online]. Available: <https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2022cv01461/575368/54/>
- [43] High Court of Justice of England and Wales, "Ayinde v. london borough of hareingey and al-haroun v. qatar national bank qpsc," 2025, [2025] EWHC 1383 (Admin), June 6, 2025. [Online]. Available: <https://www.judiciary.uk/judgments/ayinde-v-london-borough-of-hareingey-and-al-haroun-v-qatar-national-bank/>
- [44] Australian Government Department of Finance, "Foi 25-26-084: Documents concerning the targeted compliance framework final report and use of generative ai," 2025. [Online]. Available: <https://www.finance.gov.au/sites/default/files/foi-25-26-084-document-1.pdf>
- [45] British Columbia Civil Resolution Tribunal, "Moffatt v. air canada," 2024, 2024 BCCRT 149, February 14, 2024. [Online]. Available: <https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html>
- [46] NIST NCCoE, "Accelerating the adoption of software and ai agent identity and authorization (concept paper)," Tech. Rep., 2026.
- [47] R. Pak, E. Rovira, and A. McLaughlin, "Genai as complete cognitive automation for untrained users at population scale: Implications for human factors and ergonomics," *Ergonomics*, 2026. [Online]. Available: <https://doi.org/10.1080/00140139.2026.2686284>
- [48] D. Fernandes, S. Villa, S. Nicholls, O. Haavisto, D. Buschek, A. Schmidt, T. Kosch, C. Shen, and R. Welsch, "Ai makes you smarter but none the wiser: The disconnect between performance and metacognition," *Computers in Human Behavior*, vol. 175, p. 108779, 2026. [Online]. Available: <https://doi.org/10.1016/j.chb.2025.108779>
- [49] E. W. Dillon, S. Jaffe, N. Immorlica, and C. T. Stanton, "Shifting work patterns with generative ai (nber working paper no. 33795; rev. november 2025)," Tech. Rep., 2025. [Online]. Available: <https://doi.org/10.3386/w33795>

- [50] M. Afshar *et al.*, “A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. *nejm ai*, 2(12),” *NEJM AI*, 2025. [Online]. Available: <https://doi.org/10.1056/AIoa2500945>
- [51] Board of Governors of the Federal Reserve System, “Monitoring ai adoption in the u.s. economy. feds notes,” 2026.
- [52] A. Singla, A. Sukharevsky, L. Yee, M. Chui, B. Hall, and T. Balakrishnan, “The state of ai in 2025: Agents, innovation, and transformation. mckinsey & company / quantumblock,” 2025.
- [53] PwC, “Global workforce hopes and fears survey 2025,” 2025.
- [54] A. Chatterji, D. Holtz, N. Rakholia, P. Tambe, and G. Weeratunga, “How organizations use ai: Evidence from chatgpt. *arxiv:2608.12236*,” 2026.
- [55] Government Digital Service, “Microsoft 365 copilot experiment: Cross-government findings report. gov.uk,” 2025.
- [56] M. Bano, D. Zowghi, J. Whittle, L. Zhu, A. Reeson, R. Martin, and J. Parsons, “Survey insights on m365 copilot adoption. *arxiv:2412.16162*,” 2024.
- [57] A. Challapally, C. Pease, R. Raskar, and P. Chari, “The genai divide: State of ai in business 2025. mit project nanda (preliminary),” 2025.
- [58] PwC, “Ai agent survey (united states),” 2025.
- [59] U.S. Government Accountability Office, “Artificial intelligence: Irs actions needed to address skills gaps, information quality, and strategic management (gao-26-107522),” *Tech. Rep.*, 2026.
- [60] Klarna, “Klarna ai assistant handles two-thirds of customer-service chats in its first month [press release],” 2024.
- [61] S. Westby, A. Modestino, and P. Cheng, “Generative ai and the redefinition of entry-level software work (iza discussion paper no. 18723),” 2026.
- [62] E. Brynjolfsson, B. Chandar, and R. Chen, “Canaries in the coal mine? six facts about the recent employment effects of artificial intelligence. stanford digital economy lab,” 2025.
- [63] D. Acemoglu, D. Autor, J. Hazell, and P. Restrepo, “Ai and jobs: Evidence from online vacancies (nber working paper no. 28257),” *Tech. Rep.*, 2020.
- [64] S. Peng, E. Kalliamvakou, P. Cihon, and M. Demirer, “The impact of ai on developer productivity: Evidence from github copilot. *arxiv:2302.06590*,” 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2302.06590>
- [65] Y. H. Ke *et al.*, “Clinical and economic impact of a large language model in perioperative medicine: A randomized crossover trial. *npj digital medicine*, 8, 462,” *npj Digital Medicine*, 2025. [Online]. Available: <https://doi.org/10.1038/s41746-025-01858-x>
- [66] Gartner, “Gartner predicts over 40% of agentic ai projects will be canceled by end of 2027,” 2025.
- [67] “National institute of standards and technology. (n.d.). ai rmf playbook,” 2026. [Online]. Available: https://airc.nist.gov/docs/AI_RMF_Playbook.pdf
- [68] Office of Management and Budget, “M-24-10: Advancing governance, innovation, and risk management for agency use of artificial intelligence (rescinded by m-25-21),” 2024. [Online]. Available: <https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-A-dvancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf>
- [69] U.S. Bureau of Labor Statistics, “Artificial intelligence, information technology, and employment, 2024–34. the economics daily,” 2026.
- [70] “Worker adjustment and retraining notification act, 29 u.s.c. ss 2101–2109,” 2026.
- [71] F. Dell’Acqua, C. Ayoubi, H. Lifshitz-Assaf, R. Sadun, E. Mollick, L. Mollick, Y. Han, J. Goldman, H. Nair, S. Taub, and K. R. Lakhani, “The cybernetic teammate. *organization science*, 37(4),” *Organization Science*, 2026. [Online]. Available: <https://doi.org/10.1287/orsc.2025.20702>
- [72] Deloitte, “State of generative ai in the enterprise (wave 4),” 2025.
- [73] Gartner, “Gartner predicts half of companies that cut customer service staff due to ai will rehire by 2027,” 2026.
- [74] —, “Gartner survey finds 85% of service and support leaders are expanding human agent responsibilities despite expectations of mass ai layoffs,” 2026.
- [75] IDC, “Agentic ai to dominate it budget expansion [press release],” 2025.
- [76] Microsoft and LinkedIn, “2024 work trend index annual report,” 2024.
- [77] OpenAI, “The state of enterprise ai 2025,” 2025.
- [78] H.-P. Lee, A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks, and N. Wilson, “The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers,” in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–22. [Online]. Available: <https://doi.org/10.1145/3706598.3713778>
- [79] “U.s. bureau of labor statistics. quarterly census of employment and wages (qcew). industry and county employment and wage administrative file; no ai-spend item,” 2026.
- [80] “U.s. census bureau. longitudinal employer-household dynamics (lehd) quarterly workforce indicators, public-use files (r2026q1 vintage referenced). industry, geography, and firm-size aggregates; not named employers,” 2026.
- [81] Cisco Systems, Inc., “Form 8-k (item 2.05) and fiscal 2025 fourth-quarter earnings call remarks (c. robbins; s. herren) on reallocation toward ai, cloud, and cybersecurity,” 2024.
- [82] UiPath, Inc., “Form 8-k (item 2.05) (10% of 4,200). uipath, inc. (2025, march 12). form 8-k modifying restructuring cost estimates,” 2024.
- [83] Five9, Inc., “Form 10-q for the quarter ended june 30, 2025 (note on the march 31, 2025 reduction-in-force plan),” 2025.
- [84] Chegg, Inc., “Form 8-k. chegg, inc. (2025). form 10-k for the year ended december 31, 2024,” 2024.
- [85] Dropbox, Inc., “Fiscal fourth-quarter 2024 earnings exhibit 99.1 (workforce-reduction expense; dash described as still early),” 2024.
- [86] Federal News Network, “Reporting of gsa officials’ remarks on post-cut ai automation (‘million hours’) and workforce loss since fy2024,” 2026.
- [87] “U.s. department of the treasury. delivery order 2023h325f00060 (chatgpt licenses; current award value 3,398,315; up to 3,000 seats) as recorded in usaspending/fpds,” 2026.
- [88] U.S. Army, “Enterprise agreement award to palantir usg (ceiling 10 billion; not an obligation). delivery orders w56kgv25f0083 (333,333.34) and w56kgv25f0090 (1,166,666.69) for artificial intelligence platform licenses (usaspending),” 2025.
- [89] U.S. General Services Administration, “Onegov agreement with anthropic (claude). u.s. general services administration. (2025, september 3). onegov agreement with servicenow. u.s. general services administration. (2026, september 10). consumption-based onegov chatgpt follow-on,” 2025.
- [90] University of Maine System, “Planned chatgpt edu award (1.39 million / two years), as stated by the vice chancellor for finance and strategic ai integration,” 2026.
- [91] “Mississippi state university. notice of proposed sole-source purchase, claude education enterprise (not to exceed 250,000),” 2026.
- [92] University of Kent, “Find a tender award 2026/s 000-012152 to openai uk ltd (gbp 3,257,010 excluding vat),” 2026.
- [93] “U.s. bureau of labor statistics. job openings and labor turnover survey (jols). job openings levels from the march 2022 peak onward (macroeconomic context; not an ai attribution),” 2022.
- [94] U.S. Census Bureau, “Large firms with at least 20 employees biggest ai users,” 2026.
- [95] U.S. General Services Administration, “Gsa announces new partnership with openai delivering deep discount to chatgpt,” 2025.
- [96] —, “Multibillion-dollar gsa onegov agreement with microsoft brings steep discounts,” 2025.
- [97] GitHub, “Github copilot is generally available to all developers,” 2022.
- [98] Microsoft Corporation, “Form 10-k for the fiscal year ended june 30, 2025. additions to property and equipment,” 2025. [Online]. Available: <https://www.sec.gov/Archives/edgar/data/789019/000095017025100235/msft-20250630.htm>
- [99] Reuters, “Openai releasing version of chatgpt for large businesses,” 2023.
- [100] “Salesforce fy2025 q4 earnings remarks (26–27 feb. 2025) as reported in contemporaneous financial-technology press (engineer hiring freeze; 30% claimed engineering productivity),” 2025.
- [101] Bank of America earnings calls, “~2026 series): Ai/digitalization, operating leverage, and headcount mix (ops vs client-facing),” 2024.
- [102] Virginia Tech, “Pilot for microsoft 365 copilot: Outcome report. (vendor-partner pilot; excluded from independent ux synthesis.),” 2025.
- [103] —, “Chatgpt edu pilot: Findings report. (unclear independence; excluded from primary ux synthesis.),” 2025.
- [104] Department for Work and Pensions, “An evaluation of dwp’s microsoft 365 copilot trial. gov.uk,” 2026.
- [105] Department for Business and Trade, “Microsoft 365 copilot pilot: Dbt evaluation report. gov.uk,” 2025.
- [106] HM Revenue and Customs, “Evaluating the impact of microsoft copilot in hmrc. gov.uk,” 2026.
- [107] University of Washington Information Technology, “Microsoft 365 copilot pilot program report (one-pager). (unclear independence.),” 2024.
- [108] Bloomberg, “Klarna turns from ai to real-person customer service,” 2025.
- [109] Measuring agents in production, “arxiv:2512.04123,” 2025.

- [110] An empirical study of testing practices in open source AI agent frameworks and agentic applications, “arxiv:2509.19185,” 2025.
- [111] Tangent: An empirical study of testing practices for LLM-based agent applications, “arxiv:2608.08413,” 2026.
- [112] On the use of agentic coding: An empirical study of pull requests on GitHub, “arxiv:2509.14745,” 2025.
- [113] “Microsoft. control maker-provided credentials for authentication (copilot studio). microsoft learn. (architecture only),” 2026.
- [114] CIO, “Salesforce’s agentforce product maturity questioned as keybank cites weak customer traction,” 2026.